Edustavuuden kehitys kuluttajabarometrissa siirryttäessä yhdistelmätiedonkeruuseen

Heikki Hyhkö

Helsingin yliopisto
Valtiotieteellinen tiedekunta
Tilastotiede
Pro gradu -tutkielma
Huhtikuu 2020

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Valtiotieteellinen tiedekunta	Sosiaalitieteiden laitos

Tekijä — Författare — Author
Heikki Hyhkö

Työn nimi — Arbetets titel — Title

Edustavuuden kehitys kuluttajabarometrissa siirryttäessä yhdistelmätiedonkeruuseen

Oppiaine — Läroämne — Subject		
Tilastotiede		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Huhtikuu 2020	68 s.

Tiivistelmä — Referat — Abstract

Otoksen edustavuus on yksi keskeisimpiä asioita kyselytutkimusten hyvyyttä tarkasteltaessa. Edustavuutta voi mitata usealla eri tavalla. Perinteisin mittari on vastausaste. Korkea vastausaste ei kuitenkaan ole yksinään mikään tae otoksen edustavuudesta. Toimivia edustavuusmittareita on pitkään pyritty kehittämään. Yksi näistä on R-indikaattori, jota tässä tutkielmassa tarkastellaan.

Tilastokeskuksen perinteisen Kuluttajabarometrin sisältöä muutettiin toukokuussa 2019. Keskeisimmät muutokset olivat: 1) siirtyminen yhdistelmätiedonkeruuseen, 2) ikäjakauman kaventaminen 3) osan haastattelukysymyksistä vaihtuminen. Samassa yhteydessä tutkimuksen nimeksi vaihdettiin Kuluttajien luottamus.

Tämän tutkielman kannalta keskeisin mainituista muutoksista oli siirtyminen puhelinhaastatteluista yhdistelmätiedonkeruuseen. Tutkielman tarkoituksena on selvittää haastattelutavan muutoksen vaikutusta otoksen edustavuuteen. Edustavuusmittariksi valittiin R-indikaattori.

Tutkimusaineistona oli kuluttajabarometridata vuoden 2012 tammikuusta vuoden 2019 toukokuuhun. Kuluttajabarometridatan lisäksi käytössä oli Kuluttajien luottamus -tutkimuksen data neljältä ensimmäiseltä kuukaudelta toukokuusta elokuuhun 2019.

Tutkimuksen tuloksena oli, että siirtyminen yhdistelmätiedonkeruuseen ei heikentänyt otoksen edustavuutta. Toisaalta kävi kuitenkin ilmi, että R-indikaattorin saamat arvot eivät koko tutkimusperiodilla olleet valittujen hyvyysrajojen mukaan riittävän korkealla tasolla.

Toinen tarkastelluista muutoksista oli ikäjakauman kaventaminen molemmista päistä. Yläpäästä jätettiin kokonainen ikäluokka pois (75-84 v.). Alapäästä jätettiin pois osa nuorimmasta ikäluokasta (15-17 v.). Vanhin ikäluokka oli aktiivisin vastaajaryhmä ja vastaavasti nuorin ikäluokka oli passiivisin vastaamaan. Ikäjakauman kaventaminen ei kuitenkaan heikentänyt otoksen edustavuutta.

Edustavuuden kehityksen lisäksi tarkasteltiin vaihtoehtoisia edustavuusindikaattoreita ja R-indikaattorin erilaisia versioita. Suurin osa vaihtoehtoisista indikaattoreista antoi hyvin samankaltaisia tuloksia, kuin R-indikaattori. Mikään testatuista vaihtoehtoisista indikaattoreista ei osoittautunut merkittävästi helpommin tulkittavaksi kuin R-indikaattori.

Avainsanat — Nyckelord — Keywords
R-indikaattori, edustavuus, vastausaste, yhdistelmätiedonkeruu, survey

Sisältö

Luku 1

Johdanto

Tämän tutkielman tarkoituksena on selvittää otoksen edustavuuden muutosta kuluttajabarometrissa, kun tiedonkeruussa siirrytään puhelinhaastatteluista yhdistelmätiedonkeruuseen. Tilastokeskuksen perinteisen Kuluttajabarometrin nimeksi vaihdettiin toukokuussa 2019 Kuluttajien luottamus. Samassa yhteydessä tiedonkeruussa siirryttiin puhelinhaastatteluista nettikyselyn ja puhelinhaastattelun yhdistelmään. Tässä tutkielmassa on tarkoitus selvittää kuinka tämä muutos vaikuttaa otoksen edustavuuteen. Mittarina käytetään R-indikaattoria (representativeness indicator).

Pääluvussa kaksi käydään läpi aineistoon ja haastattelumenetelmiin liittyvät kysymykset. Eli aluksi käydään läpi puhelin- ja nettikyselyjen erityispiirteet ja eroavaisuudet. Tämän jälkeen selvitetään mitkä asiat muuttuvat, kun siirrytään yhdistelmätiedonkeruuseen. Seuraavaksi tarkastellaan, mitä muita muutoksia tähän kyselytutkimukseen tehtiin tässä yhteydessä. Keskeisimmät näistä muutoksista kohdistuivat kohdeperusjoukkoon, kun vanhin ikäluokka poistettiin kokonaan ja nuorimmasta ikäluokasta poistettiin alle kahdeksantoistavuotiaat. Samassa yhteydessä kyselyn varsinaisia muuttujiakin muutettiin hieman, kun EU-tasolla vaihdettiin eräiden indeksien laskuperiaatteita. Kolmas keskeinen muutos oli se, että siirryttiin käyttämään rotatoivaa paneelia, mutta tämän muutoksen vaikutusta ei tämän tutkimuksen aikavälillä pystytä havaitsemaan.

Pääluvussa kolme selvitetään aluksi, mitä tarkoitetaan edustavuudella, jonka jälkeen siirrytään tarkastelemaan itse R-indikaattorin teoriaa. Ennen kuin päästään syventymään itse edustavuuteen, täytyy kuitenkin selvittää, millainen puuttuneisuusmekanismi on kyseessä. Puuttuneisuusmekanismin selvittyä päästään tutkimaan vastausalttiutta, vastaamattomuusharhaa ja viimein itse edustavuutta. Edustavuuden tutkimiseksi pitää valita jokin edustavuusindikaattori. Tässä tutkimuksessa edustavuusmittariksi on valittu R-indikaattori. R-indikaattorin laskemiseksi täytyy aineistosta estimoida vastausalttiuden odotusarvo jollain tilastollisella mallilla. Koska kyseessä on dikotominen muuttuja (vastaa / ei vastaa), niin yksi luonnollinen vaihtoehto on logistinen regressiomalli. R-indikaattorin laskenta perustuu pohjimmiltaan muokattuun keskihajontaan, jossa tavallisen odotusarvoestimaatin sijaan käytetäänkin logistisella regressiomallilla estimoitua vastausalttiuden odotusarvoa. Tämän luvun lopuksi käydään vielä läpi joukko vaihtoehtoisia edustavuusindikaattoreita, jotta edustavuudesta saataisiin

moniulotteisempi kuva.

Pääluvussa neljä tehdään varsinaiset laskelmat Kuluttajabarometri- / Kuluttajien luottamus -aineistoon perustuen. Varsinaisen laskennan aluksi täytyy valita aineistosta tilanteeseen sopivat apumuuttujat, joiden suhteen otoksen edustavuutta tutkitaan. Kuluttajabarometriaineistoa tutkittaessa potentiaalisista apumuuttujista mallinnukseen valikoituivat ikä ja koulutustaso. Apumuuttujien valinnan jälkeen päästään estimoimaan logistisella regressiomallilla vastausalttiudet ja laskemaan R-indikaattorit luottamusväleineen. Seuraavaksi tarkastellaan laskettujen R-indikaattorien tulkintaa ja hyvyyttä aineistomme valossa. Vastausalttiuden ja R-indikaattorin muutosta ajassa tutkitaan lineaarisen regressiomallin avulla. Samassa yhteydessä lasketaan myös osittaiset R-indikaattorit ja tutkitaan myös niiden muutosta ajassa. Vertailun vuoksi lasketaan myös vaihtoehtoisten edustavuusindikaattoreiden arvot ja vertaillaan niitä R-indikaattoriin ja vastausasteeseen. Lopuksi vielä selvitetään, kuinka siirtyminen yhdistelmätiedonkeruuseen vaikutti vastausaktiivisuuteen ja R-indikaattorin arvoihin.

Pääluvussa viisi vedetään tutkimustulokset yhteen ja pohditaan mitä edustavuuden parantamisen suhteen olisi tehtävissä.

Luku 2

Aineisto

2.1 Haastattelumenetelmät

Kyselytutkimuksia tehtäessä haastattelut voidaan tehdä useilla erilaisilla menetelmillä tai niiden yhdistelmillä. Perinteisin tapa tehdä kyselytutkimuksia on PAPI (Paper and pencil interview), eli haastattelija esittää kysymyslomakkeen kysymykset suoraan haastateltavalle. Tällaisia kyselyjä käytetään usein tuotetestauksissa/-esittelyissä, joita tapahtuu mm. kauppakeskuksissa. Aiemmin kyseinen menettely oli myös tutkimuslaitosten yleisessä käytössä, kun muita vaihtoehtoja ei juurikaan ollut. Kysymykset voitiin tällöin esittää joko naamakkain tai puhelimitse. Nykyään suuret tutkimuslaitokset tekevät kyselynsä tietokoneavusteisesti. Tällaisia tietokoneavusteisia CAI (Computer-assisted interviewing) menetelmiä ovat mm. CAPI, CATI, CAWI ja CASI. Toinen perinteisempi kyselytutkimusformaatti on tietysti postikysely, eli vastaajalle postitetaan kyselylomake, jonka hän vastattuaan postittaa oheisessa kirjekuoressa kyselyn tekijälle.

CAPIssa (Computer-assisted personal interviewing) tilastohaastattelija käy haastateltavan kotona esittämässä kysymykset ja joko haastattelija tai haastateltava syöttää vastaukset kannettavalle tietokoneelle. CATIssa (Computer-assisted telephone interviewing) haastattelija soittaa haastateltavalle ja esittää kysymykset puhelimitse samalla tallentaen vastaukset tietokoneelle. CAWIssa (Computer-assisted web interviewing) ei ole haastattelijaa, vaan vastaaja vastaa tietokoneella tai älypuhelimella nettilomakkeella oleviin kysymyksiin. Ja jos paneudutaan hieman tarkemmin lyhenteeseen CAWI, niin havaitaan, että se on tautologia, mutta siitä huolimatta terminä yleisesti käytetty [de Leeuw (2005)]. Mitä taas tulee CASIin (Computer-assisted self interviewing), niin se nykyään sisältyy CAWIin (tai päinvastoin), mutta aikaisemmin tällaisia kyselytutkimuksia tehtiin myös erilaisten video- tai audionauhoitteiden avustuksella. [Laaksonen (2018)]

2.1.1 Puhelinhaastattelu

Puhelinhaastattelu on erittäin paljon käytetty haastattelumenetelmä. Puhelinhaastattelujen alkuaikoina niiden ongelmana oli puhelinverkon kattavuus, josta seurasi merkittävää alipeittävyyttä. Nykyään tämä ei ole niin suuri ongelma,

kun esimerkiksi Suomessa lähes kaikilla on puhelin. Vuonna 2013 Tilastokeskuksella oli automaattisesti saatavilla kaksi kolmasosaa puhelinliittymistä ja jopa 90 prosenttia olisi ollut saatavissa, jos tutkimuksessa on tähän tarkoitukseen varattuna riittävästi rahoitusta. [Laaksonen (2013)]

Nykyään ongelmana on enemmänkin ihmisten haluttomuus vastata tuntemattomista numeroista tuleviin puheluihin, ainakin osittain johtuen kasvaneesta puhelinmarkkinoinnista. Toinen ongelma on ”huonoon aikaan” saapuvat puhelut, kun teoriassa ihmiset ovat jatkuvasti saavutettavissa. Näitä molempia ongelmia voidaan pienentää suunnittelemalla kyselyprosessi huolellisesti. Ennen varsinaista puhelua kannattaa lähettää otokseen kuuluvalle vastaajalle sähköposti tai kirje, jossa kerrotaan, että tällainen kysely on tulossa ja vastaaja on osunut otokseen. Samalla kannattaa selittää mistä tutkimuksessa on kyse ja miksi on oleellista, että vastaaja osallistuu kyselyyn.

Puhelinhaastattelussa on muutamia ilmeisiä hyötyjä suhteessa itse täytettävään lomakkeeseen. Yksi selkeimmistä on epäselvien ja epäkelpojen vastausten väheneminen, kun virheellisesti muotoiltuja vastauksia ei päädy lomakkeelle. Esimerkiksi lukumuodot ja mittayksiköt tulevat tällöin oikein kirjatuiksi. Myöskään kysymyksiä tai sivuja ei tule epähuomiossa ohitettua. Täytetty lomakekaan ei jää tällöin palauttamatta, kun se on haastattelijan hallussa. Haastateltavan kannalta haastattelijasta on myös hyötyä, kun epäselvissä tapauksissa haastateltava voi kysyä tarkentavia kysymyksiä. Ja vastaavasti haastattelija voi tarvittaessa täsmentää kysymyksiä, jos vaikuttaa siltä, että vastaaja on jollain lailla epätietoinen. Haastattelija hoitaa myös, kyselyyn mahdollisesti sisältyvien vain osaa vastaajista koskevien kysymysten sivuuttamiset, ilman tarvetta vaivata vastaajaa. Haastattelijan läsnäolo mahdollistaa ”eos” -tyyppisten vastausvaihtoehtojen esittämisen vasta tarvittaessa, joka ei itse täytettävällä paperilomakkeella ole mahdollista. (Edellä käytetty lyhennys ”eos” muodostuu sanoista ”ei osaa sanoa”.) Eikä sovi unohtaa, että haastattelijalla on myös mahdollisuus vakuuttaa haastateltava kyselyyn vastaamisen tärkeydestä. [Dillman (2014)]

Tietokoneavusteisuudesta saavutetaan muitakin hyötyjä. Jos ja kun tiedossa on otokseen osuneen puhelinnumero, niin järjestelmä hoitaa numerovalinnan ja mahdollistaa uusintasoittojen ajoittamisen vastaajalle soveliaampaan aikaan. Järjestelmä tallentaa myös ns. paradataa, eli puhelun keston, puhelun ajankohdan ja soittoyritysten määrän. Haastattelija pystyy myös lisäämään omat kommenttinsa järjestelmään mahdollisen kadon syystä ja muista vastaavista aiheista. Järjestelmä tukee myös haastattelijaa, kun kysymykset vastausvaihtoehtoineen voi lukea suoraan tietokoneen näytöltä. Tietokone helpottaa myös haastattelijan toimintaa, kun sen avulla pystyy tarvittaessa ohittamaan vastaajaa koskemattomat lisäkysymykset. Järjestelmästä nähdään myös taustatiedot haastateltavasta, joiden perusteella haastattelija voi mm. varmistaa, että kyse on oikeasta henkilöstä. [Dillman (2014)]

2.1.2 Nettikysely

Tietokoneavusteinen nettikysely on perinteisen kirjekyselyn kehittyneempi versio. Tilastokeskustyyppisessä viitekehyksessä nettikyselyjen vastaajien valinta ei merkittävästi poikkeaa puhelinhaastattelujen vastaajien valinnasta, sillä vastaajien

valintaan käytetään sunnilleen samaa menettelyä. Suurella osalla täysi-ikäisistä on olemassa sähköpostiosoite ja jos tämä ei ole kyselyn tekijän saatavissa, niin otokseen osuneelle lähetetään kirje, jossa heitä pyydetään vastaamaan kyselyyn. Termillä nettikysely viitataan tässä tutkielmassa aitoihin otantatutkimuksiin, eikä netistä pilvin pimein löytyviin itsevalikoituviin näytteisiin perustuviin kyselyihin.

Ilmeinen hyöty nettikyselyissä verrattuna puhelinhaastatteluun on se, että haastateltava voi vastata kyselyyn silloin, kun se hänelle parhaiten sopii. Vastaavasti ilmeinen haitta on se, että ei ole haastattelijaa vakuuttamassa vastaamisen tärkeydestä, joten kyselyn sijoittaminen ”roskakoriin” on valitettavan helppoa. Tämän vuoksi onkin välttämätöntä pyrkiä vakuuttamaan vastaaja vastaamisen tärkeydestä.

Tarkentavien kysymysten esittäminen normaalissa nettikyselyssä ei ole mahdollista, joten kyselyssä on annettava kaikki mahdollinen tieto, riippumatta siitä tarvitseeko yksittäinen vastaaja sitä vai ei. Tämä mahdollisesti vaikuttaa vastaamiseen, sillä haastattelijan esittäessä kysymykset vastaamisen kannalta kaikkea informaatioita ei tarvitse esittää, jos se ei vastaamisen kannalta ole oleellista. Kuten puhelinhaastattelun yhteydessä aiemmin mainittiin, niin nettilomakkeessa ”eos” -tyyppisten vastausvaihtoehtojen olisi oltava valittavissa. Tai ei se toki täysin välttämätöntä ole, jos vastaaja halutaan pakottaa antamaan mielipide kyseiseen kysymykseen. Tällaista pakottamista ei kuitenkaan kannata tehdä kuin äärimmäisissä tilanteissa, joissa mielipiteen saaminen kyseiseen kysymykseen on tutkimuksen kannalta välttämätöntä. Sopivan vastausvaihtoehdon puuttuminen pakottaa vastaajan vastaamaan omien mielipiteidensä vastaisesti tai vaihtoehtoisesti täysin satunnaisesti, jos kysymys ei syystä tai toisesta kosketa häntä tai hänen elinpiiriään. Tähän samaan aihepiiriin liittyy myös kysymys siitä, että tuleeko kaikkiin kysymyksiin vastata, jotta pääsee etenemään kyselyssä. Ja vielä kolmas saman tyyppinen kysymys on mahdollisuus peruuttaa, eli palata korjaamaan mahdollinen virheellinen vastaus. Edellä mainitut kolme ongelmaa voivat turhauttaa vastaajaa ja saattavat johtaa kyselyn keskeyttämiseen tai vastaamismotivaation laskuun, joka saattaa puolestaan vaikuttaa myöhempiin vastauksiin. Kumpikaan vaihtoehto ei ole hyväksi tekeillä olevalle tutkimukselle. [Dillman (2014)]

Teknisenä haasteena nettikyselyissä on se, että vastaaminen pitäisi saada sujuvaksi kolmella erilaisella alustalla, eli älypuhelimella, tietokoneella ja niiden välimuodolla taulutietokoneella. Puhelimen tapauksessa ongelma on mahdollisesti pitkien kysymysten mahduttaminen pienelle näytölle ja tekstimuotoisten vastausten kirjoittaminen. Toinen tekninen tai oikeastaan ohjelmallinen kysymys on vastausten toimittaminen. Vaihtoehdothan ovat lähettää vastaus jokaisen kysymyksen jälkeen, jokaisen kysymyssarjan jälkeen tai vasta koko kyselyn lopussa. Nykyisten hyvien nettiyhteyksien aikana reaaliaikainen vastausten siirtyminen palvelimelle toimii hyvin ja mahdollistaa myös tarvittaessa vanhoihin vastauksiin palaamisen. Tämä kysymyshän on oleellinen vain siinä tapauksessa, että osittaisetkin vastaukset ovat tutkimuksen kannalta kiinnostavia. Jos näin on, niin tutkimuksen kannalta oleellisimmat kysymykset kannatta laittaa kyselyn alkuosaan. [Dillman (2014)]

2.1.3 Yhdistelmätiedonkeruu

Yhdistelmätiedonkeruulla (Mixed mode survey) voidaan viitata kaikkiin tiedonkeruumenetelmiin, joissa käytetään useampaa kuin yhtä menetelmää joko itse tiedon keruussa tai yhteydenotossa. Tässä tutkielmassa keskitytään ensin mainittuun ja erityisesti puhelinhaastattelun ja nettikyselyn yhdistelmään. Keskeisimmät syyt yhdistelmätiedonkeruuseen siirtymiseen liittyvät yleensä tarpeeseen ehkäistä vastauskadon kasvua tai vaihtoehtoisesti haluun parantaa otoksen edustavuutta. Mainitut vaihtoehdot eivät toki ole tosiaan poissulkevia, vaan pikemminkin toisiaan tukevia. Kolmas merkittävä syy on kulujen kurissa pitäminen. Siirtymävaiheessa kulutaso toki nousee, mutta myöhemmin nettikyselyn halvempi hinta painaa kulutason alemmas. Myös mahdollisuus säästää aikaa voi olla yhtenä motivaationa yhdistelmätiedonkeruuseen siirryttäessä. Puhelinhaastattelusta nettikyselyyn laajennettaessa ei valitettavasti saavutettane ajallista hyötyä vain ennemminkin päinvastoin. Puhelinhaastattelussa vastaus saadaan välittömästi, teoriassa nettikysely toimii samalla tavalla, mutta saattaa vaatia muistutuksia. Toki myös uudelleensoitto on mahdollinen, jos hetki ei ole otollinen, mutta todennäköisemmin kieltäytyminen haastattelusta on välitön. Valitulla tiedonkeruumenetelmällä on myös vaikutusta tutkimuksesta saataviin tuloksiin. Nämä vaikutukset eivät aina välttämättä ole merkittäviä, mutta ne tulee kuitenkin ottaa huomioon jo tutkimusta suunniteltaessa. [Dillman (2014)]

Puhelinhaastattelun ja nettikyselyn yksi keskeisimmistä eroista on haastattelijan läsnäolo. Edellisissä luvuissa käsittelemättä jätettiin kysymys haastattelijan läsnäolon sosiaalisesta puolesta. Ihmisillä on havaittu luontainen haluttomuus välittää ikäviä uutisia, kun taas iloisten uutisten välittäminen tuntuu mukavalta. Jo muinaiset kreikkalaiset (mm. Sofokles sekä myöhemmin keskiajalla Shakespeare) havaitsivat tämän inhimillisen piirteen, jonka pimeä puoli tiivistyy fraasissa: ”Älä tapa viestintuojaa!”. Toisin sanoen ihmiset keskimäärin haluavat luontaisesti miellyttää keskustelukumppaniaan, vaikka tämä olisikin itselle tuntematon tilastohaastattelija. Tämä ns. MUM-efekti [Rosen (1970)] vaikuttaa haastattelutilanteissa siten, että ihminen alitajuisesti suosii konventionaalisia vastauksia ja välttelee äärimmäisiä vaihtoehtoja. Tästä seuraa, että nettikyselyn ja puhelinhaastattelun eksaktisti samanlaisetkin kysymykset, saattavat tuottaa hieman erilaisia vastauksia.

Toinen jo aiemmin mainittu aihe liittyy niin sanottuihin neutraaleihin vastausvaihtoehtoihin, kuten ”eos” ja ”ei mielipidettä”. Kuten aiemmin mainittiin, niin tällaiset vaihtoehdot voidaan puhelinhaastattelussa jättää antamatta ensi vaiheessa. Vasta, jos vastaaja ei osaa valita, hänelle tarjotaan tällainen vaihtoehto. Nettikyselyssä kaikki vastausvaihtoehdot ovat heti saatavilla ja tarjoavat helpon tavan olla rasittamatta mieltään tarkempaa pohdintaa vaativissa kysymyksissä. Puhelinhaastattelussa myös vastaajan muistille asetetaan isompia haasteita, koska hän ei pysty palaamaan jo lueteltuihin vaihtoehtoihin. Eli samalla, kun vastaajan tulisi miettiä varsinaista kysymystä, hänen täytyy myös muistella vastausvaihtoehtoja. Nämä asiat kannattaa ottaa huomioon yhdistelmätiedonkeruuta suunniteltaessa ja myös myöhemmin tuloksia tulkittaessa.

Dillmannin (2014) kirjassa on yhteensä 184 "ohjenuoraa"(guideline), hyvän kyselyn aikaansaamiseksi. Ohjenuorat on numeroitu kaksiportaisesti mukaillen kirjan

lukuja. Varsinaisesti yhdistelmätiedonkeruuta käsitteleviä lukuja/ohjenuoria on 23 ja ne jakautuvat seuraaviin aihealueisiin:

- Erojen minimointi kyselytyyppien välillä (ohjenuorat: 11.1–11.8)
- Erilaisten kontaktointitapojen hyödyntäminen (ohjenuorat: 11.9–11.13)
- Erilaisten kyselytyyppien tarjoaminen (ohjenuorat: 11.14–11.19)
- Yhdistelmätiedonkeruun testaaminen (ohjenuorat: 11.20–11.23)

Ensimmäiset luvut käsittelevät kyselyjen laadintaa siten, että tiedot olisivat mahdollisimman yhteensopivia. Toisessa joukossa tavoitellaan mahdollisimman laajaa saavutettavuutta käyttäen erilaisia viestintävälineitä, kuten puhelin, posti ja sähköposti. Kolmannessa ryhmässä tutkaillaan kyselytyyppien vaikutusta vastausasteisiin. Viimeisessä osiossa tarkastellaan tapoja, joilla voidaan tarkkailla eri kyselytyyppien tuottamien aineistojen yhteensopivuutta.

Tämän tutkielman kannalta mielenkiintoisinta on kolmannen osion havainto, että kyselytyyppien vastaajaprofiilit poikkeavat toisistaan ja samalla osittain täydentävät toisiaan. Mutta samalla korostetaan, että mitään huikeaa parannusta vastauskadon vähenemiseen ei kannata odottaa. Toinen mielenkiintoinen havainto on, että vaihtoehtoisten vastaustapojen tarjoaminen saattaa jossain tapauksessa jopa lisätä kieltäytymisiä, kun tulee yksi päätösvaihe lisää.

Kannattaa huomioida, että yhdistelmätiedonkeruuesimerkkejä tarkasteltaessa kohdatut haasteet olivat olemassa jo alkuperäisissä kyselytutkimuksissa, mutta ne nousivat esiin vasta, kun siirryttiin yhdistelmätiedonkeruuseen. Toisena huomiona mainittakoon, että vertailtavuus taaksepäin saattaa heiketä siirtymävaiheessa, mutta ajan kuluessa lopullinen tulos on kuitenkin edustavampi.

Huomautettakoon vielä lopuksi, että yllättävän suuri osa Dillmannin kirjan ehdotuksista vastauskadon korjaamiseksi, pitää sisällään rahallisia tai muunlaisia korvauksia vastaamisesta. Tämä ei ole ainakaan tällä hetkellä mahdollista tilastokeskustyyppisessä viitekehyksessä.

2.2 Kulutustutkimus

Sana barometri tarkoittaa varsinaisesti ilmapuntaria, mutta sitä käytetään myös erilaisista tilastollisista mittareista puhuttaessa. ”Barometri tarkoittaa asenteiden ja mielipideilmaston mittausta tai mittauksen tulosta. Barometri tehdään kyselytutkimuksena. Toistuvia kyselytutkimuksia ovat esimerkiksi kuluttaja-, suhdanne- ja eurobarometri.” [Moilanen (2011)]. Tilastollisessa käytössä barometri ilmaisee yleensä jonkin asian ajasta riippuvaa kehitystä. Hyvin usein nämä barometrit mittaavat taloudellista tilannetta.

Ensimmäinen barometrityyppinen kysely tehtiin Yhdysvalloissa vuonna 1939. Kulutustutkimuksen varsinaisena alullepanijana voidaan pitää unkarilaissyntyistä George Kantonaa, joka toimi Michiganin yliopistossa taloustieteen ja psykologian professorina noihin aikoihin. Eurobarometri aloitettiin vuonna 1972 ja Tilastokeskus aloitti kuluttajabarometrinsa vuonna 1987. Aluksi kyselyjä oli vain kaksi vuodessa.

Vuonna 1992 siirryttiin neljännesvuosittaiseen kyselyyn. Lokakuusta 1995 asti tiedot on kerätty kuukausittain EU-maiden harmonisoidun mallin mukaan. Vuoteen 2000 asti kyse oli rotatoivasta paneelista, jossa sama henkilö vastasi kolmesti puolen vuoden välein. Vuodesta 2000 alkaen käytettiin kokonaan vaihtuvaa henkilöotosta. Vuoden 2019 toukokuussa siirryttiin jälleen rotatoivaan paneeliin samalla, kun siirryttiin yhdistelmätiedonkeruuseen. [SVT (2017)]

Tässä tutkielmassa tutkimusaineistona on entisen Kuluttajabarometrin ja nykyisen Kuluttajien luottamus -tutkimuksen kyselyaineistot vuoden 2012 tammikuusta vuoden 2019 elokuuhun asti. Sekä Kuluttajabarometrissa että Kuluttajien luottamus -tutkimuksessa vastauskato huomioidaan painottamalla ositteet vastaamaan perusjoukon jakaumaa valittujen apumuuttujien suhteen.

2.2.1 Kuluttajabarometri

Edustavassa otoksessa valittujen apumuuttujien jakaumat ovat otoksessa samat kuin perusjoukossa. Kuluttajabarometrissa näiksi apumuuttujiksi on valittu suuralue (pääkaupunkiseutu, Etelä-Suomi, Länsi-Suomi sekä Itä- ja Pohjois-Suomi), ikäluokka (15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84) ja sukupuoli. Eli tässä tapauksessa halutaan, että otoksessa on samassa suhteessa näihin ositteisiin (yht. 70) kuuluvia alkioita kuin perusjoukossa. Näillä valinnoilla taataan otoksen alueellinen, ikäryhmittäinen ja sukupuolittainen edustavuus. Kuluttajabarometrissa käytetään maantieteellisesti itsepainottavaa systemaattista otantaa. Otoskoko tutkimuksessa on 2350 henkilöä.

Kyselyaineisto painotetaan vastaamaan perusjoukon jakaumaa valittujen apumuuttujien suhteen käyttämällä henkilöpainoja. Henkilöpainot lasketaan Calmar-kalibrointimenetelmällä, siten että ne vastaavat perusjoukon reunajakaumia valittujen apumuuttujien suhteen.

Ikäpyramidi ei ole todellisuutta nykypäivän Suomessa, todellinen väestörakenteen muoto vastaa ennemminkin alaspäin kapenevaa ikäpagodaa. Toisin sanoen katto levenee odotetusti, mutta muilta osin yleislinja on kapeneva vaikkakin hieman rönsyilevä ikäluokkien nuorentuessa. Edellä mainittu selittää sen, miksi keskimmäiset ikäluokat (25–74) ovat otoksessa suunnilleen samankokoisia.

Kuluttajabarometrissa vastaajien näkemyksiä kysymysajankohdan taloustilanteesta kartoitetaan seuraavilla kysymyksillä:

1. Mikä on kotitaloutenne taloustilanne nyt verrattuna tilanteeseen 12 kuukautta sitten?
2. Mikä on kotitaloutenne taloustilanne 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
3. Mikä on Suomen taloudellinen tila nyt verrattuna tilanteeseen 12 kuukautta sitten?
4. Mikä on Suomen taloudellinen tila 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
5. Millä tasolla kuluttajahinnat ovat nyt verrattuna tilanteeseen 12 kuukautta sitten?
6. Montako prosenttia kuluttajahinnat ovat muuttuneet viimeisen 12 kuukauden aikana?
7. Miten kuluttajahinnat muuttuvat seuraavan 12 kuukauden aikana?

8. Montako prosenttia kuluttajahinnat muuttuvat seuraavan 12 kuukauden aikana?
9. Miten työttömyystilanne muuttuu Suomessa seuraavan 12 kuukauden aikana?
10. Onko nyt yleisesti ottaen hyvä vai huono aika hankkia kestokulutustavaroita?
11. Käytättekö kestokulutustavaroiden hankintaan rahaa seuraavan 12 kuukauden aikana enemmän, yhtä paljon vai vähemmän kuin viimeisen 12 kuukauden aikana?
12. Onko nyt yleisesti ottaen hyvä vai huono aika säästää?
13. Kuinka todennäköisesti kotitaloutenne pystyy säästämään rahaa seuraavan 12 kuukauden aikana?
14. Mikä on kotitaloutenne rahatilanne tällä hetkellä?
15. Kuinka todennäköisesti kotitaloutenne ostaa henkilöauton 12 kuukauden sisällä?
16. Aikooko kotitaloutenne ostaa tai rakentaa asunnon 12 kuukauden sisällä?
17. Kuinka todennäköisesti kotitaloutenne käyttää suuren summan rahaa kodin perusparannuksiin 12 kuukauden sisällä?

Monivalintakysymyksissä käytetään viisiluokkaista asteikkoa: paljon enemmän - hieman enemmän - yhtä paljon - hieman vähemmän - paljon vähemmän. Toki varsinaiset vaihtoehdot on sovitettu kyseiseen kysymykseen sopiviksi ja lisäksi on mahdollista vastata eos (ei osaa sanoa). Kysymyksessä 12 on kolme vastausvaihtoehtoa (edullinen - ei kumpikaan - epäedullinen) ja inflaatiokysymyksiin vastataan prosentteina.

Näistä vastauksista lasketaan painottamalla yksi saldoluku ja kolme indikaattoria. Saldoluku vastaa EU:n tasapainolukua (balance figure). Vastauksista lasketaan saldoluku, jossa äärimmäiset vastaukset saavat arvon 1 tai -1. Maltillisemmille vaihtoehdoille annetaan arvo 0,5 tai -0,5. Keskimmäinen vaihtoehto ja eos jätetään huomiotta. Kysymyksistä 1, 2, 4, ja 11 lasketaan kuluttajien luottamusindikaattori, joka vastaa EU-tason kuluttajien luottamusindikaattoria (Consumer confidence indicator). Samoja muuttujia käytetään EU-tasolla myös osana yleisemmän talouden ilmapiiriä kuvaavan indikaattorin ESI:n (Economic Sentiment Indicator) laskennassa. EU-tasolla tehdään aikatasoitus DAINTIES-metodilla. [DG-ECFIN (2019)]

EU-kysymysten lisäksi kysytään Suomessa myös yhdeksän Tilastokeskuksen omaa kysymystä. Mainittujen varsinaisten tutkimuskysymysten lisäksi esitetään taustakysymyksiä, joilla luonnehditaan vastaajaa ja hänen kotitalouttaan. Otosaineistossa ovat valmiina vastaajan perustiedot: ikä, sukupuoli ja asuinkunta. Tutkintorekisteristä saadaan lisäksi tiedot vastaajan koulutustasosta. [SVT (2019)]

2.2.2 Kuluttajien luottamus

Kuluttajien luottamus -tilasto on keskeisiltä osiltaan hyvin samankaltainen kuin sitä edeltänyt Kuluttajabarometri. Joitain muutoksia on silti tehty. Useita kysymyksiä on muutettu, koska EU-tasolla uudistettiin indeksejä. Siirtyminen puhelinhaastattelusta yhdistelmätiedonkeruuseen aiheutti myös joitain muutoksia. Vastaajien ikähaitaria

myös kavennettiin sekä ylä- että alapäästä. Lisäksi palattiin rotatoivaan paneeliin, jossa samat vastaajat vastaavat kyselyyn kolmen kuukauden kuluttua uudestaan. Tämä lisännee hieman vastausaktiivisuutta, sillä jo kerran kyselyyn vastanneet vastaajat vastaavat todennäköisemmin, kuin täysin uudet vastaajat.

Kuluttajien luottamus -tilastossa käytetyt apumuuttujat ovat: suuralue (pääkaupunkiseutu, Etelä-Suomi, Länsi-Suomi sekä Itä- ja Pohjois-Suomi), ikäluokka (18–24, 25–34, 35–44, 45–54, 55–64, 65–74), sukupuoli ja uutena muuttujana koulutus. Ikäluokkien vähentämisen seurauksena otoskoko on pudotettu 2200 henkilöön. Painotus tehdään Calmar-kalibrointimenetelmällä, kuten tehtiin jo kuluttajabarometrin aikana. Vuoden 2018 alusta lähtien painotuksessa on huomioitu myös vastaajan koulutustaso. [SVT (2019)]

Kuluttajien luottamus -tilastossa kysymykset käsittelevät pääosin vastaajan omaa taloudellista tilannetta. Kuluttajabarometrissa oltiin kiinnostuneita kotitalouden taloudellisesta tilanteesta. Kuluttajien luottamus -kyselyn kysymykset ovat seuraavat:

B1 Millainen on mielestäsi oma taloudellinen tilanteesi nyt, verrattuna tilanteeseen 12 kuukautta sitten?

B2 Entä millaisen arvioit sen olevan 12 kuukauden kuluttua, verrattuna tilanteeseen nyt?

B3 Millainen on sinun mielestäsi Suomen taloudellinen tilanne nyt, verrattuna tilanteeseen 12 kuukautta sitten?

B4 Entä millaisen arvioit sen olevan 12 kuukauden kuluttua, verrattuna tilanteeseen nyt?

B5 Kuinka paljon arvioit hintojen [nousseen / laskeneen] prosentteina viimeisen 12 kuukauden aikana? Voit antaa luvun yhden desimaalin tarkkuudella.

B6 Kuinka paljon arvioit hintojen [nousevan / laskevan] prosentteina seuraavan 12 kuukauden aikana? Voit antaa luvun yhden desimaalin tarkkuudella.

B7 Miten arvioit työttömien määrän muuttuvan Suomessa? Arveletko, että työttömiä on 12 kuukauden päästä:

B8 Onko työttömyyden tai lomautuksen uhka omalla kohdallasi viimeisen 12 kuukauden aikana mielestäsi:

C1 Jos ajattelet yleistä taloudellista tilannetta Suomessa, niin millainen aika mielestäsi nyt on ostaa kestokulutustavaroita?

C2 Jos ajattelet yleistä taloudellista tilannetta Suomessa, niin millainen aika mielestäsi nyt on säästää?

C3 Jos ajattelet taas yleistä taloudellista tilannetta, niin millainen aika mielestäsi nyt on ottaa lainaa?

D1 Mikä seuraavista vaihtoehdoista kuvaa parhaiten omaa rahatilannettasi tällä hetkellä?

D2 Kuinka todennäköistä on, että säästät rahaa seuraavan 12 kuukauden aikana?

D3 Aiotko ottaa lainaa seuraavan 12 kuukauden aikana?

E1 Verrattuna edelliseen 12 kuukauteen, miten aiot käyttää rahaa kestokulutustavaroiden hankintaan seuraavan 12 kuukauden aikana?

E2 Kuinka todennäköistä on, että käytät rahaa henkilöauton ostoon seuraavan 12 kuukauden aikana?

E3 Aiotko käyttää rahaa asunnon ostoon tai talon rakentamiseen seuraavan 12 kuukauden aikana?

E4 Kuinka todennäköistä on, että käytät suuren summan rahaa asunnon korjauksiin tai parannuksiin seuraavan 12 kuukauden aikana?

"Kuluttajien luottamusindikaattori (A1) tiivistää kuluttajien näkemykset taloudesta. Luottamusindikaattori on neljän saldoluvun aritmeettinen keskiarvo: kuluttajan oma talous nyt (B1), kuluttajan oma talous 12 kuukauden kuluttua (B2), Suomen talous 12 kuukauden kuluttua (B4) ja kuluttajan rahankäyttö kestotavaroihin seuraavan 12 kuukauden aikana verrattuna edelliseen 12 kuukauteen (E1). Tämä uusi vuonna 2019 käyttöön otettu luottamusindikaattori on Euroopan komission talous- ja rahoitusyksikön (DG ECFIN) käyttämä ja suosittelema."[SVT (2019)]

Luku 3

Teoria

3.1 Edustavuus

Edustava otos vastaa mahdollisimman hyvin perusjoukkoa. Täydellinen edustavuus on tietenkin mahdotonta, jollei kyse ole kokonaistutkimuksesta tai hyvin kompaktista populaatiosta. Yleensä pyritäänkin siihen, että otos edustaisi mahdollisimman hyvin perusjoukkoa, valittujen apumuuttujien suhteen.

Jos ajatellaan vaikkapa puolueiden kansallisia kannatusmittauksia, niin vaalijärjestelmän vuoksi tuntuisi järkevältä huomioida vaalipiirijako otantaa suunniteltaessa. Vaalipiirin lisäksi kiinnostavia apumuuttujia olisivat tässä tilanteessa ainakin äänestäjän ikä, sukupuoli ja äänestyskäyttäytyminen edellisissä vaaleissa. Muita mahdollisesti kiinnostavia apumuuttujia olisivat varmaankin tulotaso, työmarkkinastatus, koulutusaste ja siviilisääty. Valittujen luokittelevien muuttujien määrä pitää kuitenkin pitää rajattuna. Käytännössä saatavissa olevat taustatiedot yleensä rajaavatkin valinnat muutamaan kyselyn kannalta keskeisimpään apumuuttujaan. Valittu otoskoko vaikuttaa myös siihen, kuinka moneen luokkaan otos voidaan järkevästi osittaa, siten että jokaiseen ositteeseen jää riittävä määrä alkioita. Puolueiden kannatuskyselyissä pitänee siis rajoittua muutamaan oleellisimpaan taustatekijään, eli kyseeseen tulisivat varmaankin ikä, sukupuoli ja asuinkunta.

Vastauskadolla on luonnollisesti varsin merkittävä vaikutus edustavuuteen. Edustavuuden kannalta vastauskato voidaan jakaa kolmeen ryhmään täysin satunnaiseen, satunnaiseen suhteessa johonkin muuttujaryhmään ja ei-satunnaiseen. Näitä puuttuneisuuden ryhmiä käsitellään hieman tarkemmin seuraavassa luvussa.

3.1.1 Puuttuneisuus

Donald B. Rubin on määritellyt puuttuneisuudelle kolme tasoa:

1. Täysin satunnainen puuttuneisuus (MCAR)
2. Satunnainen puuttuneisuus suhteessa muuttujaryhmään (MAR)
3. Aineistosta riippuva puuttuneisuus (MNAR)

Rubin ei alkuperäisessä artikkelissaan käyttänyt vielä yllä mainittuja myöhemmin vakiintuneita nimityksiä. Esimerkiksi MCAR (Missing Completely at Random) määrittyy Rubinin alkuperäisin termein, kun MAR (Missing at Random) ja OAR (Observed at Random) ovat molemmat yhtä aikaa voimassa. [Rubin (1976)]

Täysin satunnainen puuttuvuus MCAR (Missing Completely at Random) on mahdollista lähinnä erilaisissa koeasetelmissa, missä koko perusjoukko on aidosti tutkijan hallussa. Täysin satunnainen puuttuneisuus määritellään kaavalla:

$$f\left(M|Y,\phi\right) = f\left(M|\phi\right) \quad kaikilla\, Y, \phi \tag{3.1}$$

Kaavassa Y on täysi aineisto. M on puuttuneisuusindikaattori, joka saa arvon 1, jos havainto puuttuu ja arvon 0, jos havainto on aineistossa. Puuttuneisuusmekanismin määrää M:n ehdollinen jakauma Y:n suhteen, jossa ϕ kuvaa tuntemattomia parametreja. Tässä tilanteessa puuttuneisuus on riippumatonta aineistosta. Otetaan käytännön esimerkki, eli vaikkapa aiemmin mainittu puoluekannatuskysely. Jos todennäköisyys sille, että muuttujan puoluekannatus arvo puuttuu, on sama kaikilla henkilöillä riippumatta heidän puoluekannastaan ja tulotasostaan, niin puuttuneisuus on täysin satunnaista.[Little (2002)]

Satunnainen puuttuneisuus suhteessa valittuun muuttujaryhmään MAR (Missing at Random) määrittyy kaavalla:

$$f\left(M|Y\right) = f\left(M|Y_o,\phi\right) \quad kaikilla\, Y_m, \phi \tag{3.2}$$

Tässä tilanteessa puuttuneisuus on riippumatonta puuttuvista havainnoista Y_m, mutta se saa riippua havaituista havainnoista Y_o. Ja aiemman esimerkin mukaisesti: Jos todennäköisyys sille, että muuttujan puoluekannatus arvo puuttuu, vaihtelee tulotason suhteen, mutta ei puoluekannan suhteen saman tuloluokan sisällä, niin puuttuneisuus on satunnaista suhteessa muuttujaryhmään.[Little (2002)]

Kolmannessa tilanteessa M:n jakauma riippuu puuttuvista havainnoista, eli puuttuneisuus ei ole satunnaista MNAR (Missing not at Random). Ja jälleen esimerkin mukaan: Jos todennäköisyys sille, että muuttujan puoluekanta arvo puuttuu, vaihtelee tuloluokan sisällä, niin puuttuneisuus on ei-satunnaista. Toisinaan tästä käytetään myös lyhennystä NMAR (Not Missing at Random).[Little (2002)]

3.1.2 Vastausalttius ja edustavuus

Alun perin vastausalttiuden käsitteen loivat Paul Rosenbaum ja Donald Rubin artikkelissaan *"The Central Role of the Propensity Score in Observational Studies for Causal Effects"*. Vastausalttius ρ_i on todennäköisyys, että vastaaja i vastaa, kun hän kuuluu otokseen. Määritelmässä 3.3 r_i on vastausindikaattori, joka saa arvokseen yksi, jos vastaaja vastaa ja nolla muulloin. Vastaavasti s_i on otosindikaattori, joka saa arvokseen yksi, jos havainto kuuluu otokseen ja nolla muulloin. [Rosenbaum (1983)]

Määritelmä 3.3. *Vastausalttius*

$$\rho_i = P(r_i = 1|s_i = 1)$$

Vastanneiden osajoukko on vahvasti edustava (määritelmä 3.4) suhteessa otokseen, jos vastausalttius ρ_i on sama kaikille vastaajille perusjoukossa ja vastaus on riippumaton muista vastauksista perusjoukossa.

Määritelmä 3.4. *Vahva edustavuus*

$$\rho_i = P(r_i = 1 | s_i = 1) = \rho, \forall i$$

Jos edellä mainittu olisi voimassa kaikille vastauksille, niin puuttuneisuus olisi täysin satunnaista (MCAR). Valitettavasti tämä ei ole käytännön tilanteissa mitenkään todennettavissa. Edellä mainitun vuoksi tarvitaan käytännöllisempi määritys hieman heikommalle edustavuudelle. [Rosenbaum (1983)]

Vastanneiden osajoukko on heikosti edustava (määritelmä 3.5) suhteessa otokseen, jos vastausalttius ρ_i on keskimäärin sama valittujen apumuuttujien määrittelemissä osajoukoissa h_i.

Määritelmä 3.5. *Heikko edustavuus*

$$\overline{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \textit{kaikille } h\text{=}1,2,\ldots,H$$

Edellä mainitussa tilanteessa puuttuneisuus on vastanneiden osajoukossa täysin satunnaista suhteessa valittuihin luokitteleviin muuttujiin (MAR). [Rosenbaum (1983)]

3.1.3 Vastaamattomuusharha

Otoksen edustavuus on harvoin täydellistä, sillä lähes aina otoksesta puuttuu enemmän tai vähemmän satunnaisia vastaajia. Tämän puutteen vuoksi estimaateissa esiintyy harhaa, joka johtuu vastaajien puuttumisesta. Tätä harhaa kutsutaan vastaamattomuusharhaksi (nonresponse bias). Teoreettisesti tämä harha määrittyy havaittujen havaintojen keskiarvon $\overline{y}_o$ ja aidon keskiarvon $\overline{y}_s$ erotuksena. Harhan yleinen määritelmä löytyy luvusta 3.2.6 (määritelmä 3.21).

Määritelmä 3.6. *Vastaamattomuusharha*

$$B(\overline{y}) = \overline{y}_o - \overline{y}_s = \frac{m_s}{n_s} (\overline{y}_o - \overline{y}_m)$$

Määritelmästä näemme, että toinen tapa laskea vastaamattomuusharha olisi kertoa vastaamattomuusaste (m_s/n_s) vastanneiden (y_o) ja vastaamattomien (y_m) keskiarvojen erotuksella. [Groves (2009)] Valitettavasti molemmissa laskentatavoissa tarvitaan tietoja, joita otoksesta ei ole saatavilla ($\overline{y}_s$ ja $\overline{y}_m$), joten määritelmän

mukaisella kaavalla ei käytännön tilanteissa päästä vastaamattomuusharhaa laskemaan. [Groves (2006)]

Täytyy siis lähestyä ongelmaa toisesta suunnasta. Kuten tiedetään, on otoskeskiarvo perusjoukon keskiarvon harhaton estimaattori. Myöhempiä vaiheita varten kerrotaan ja jaetaan muuttuja vielä nollasta poikkeavalla vakiolla $\overline{\rho}$.

$$\overline{Y} = \frac{1}{N}\sum_{h=1}^{N} Y_h = \frac{1}{N}\sum_{h=1}^{N} \frac{\overline{\rho} Y_h}{\overline{\rho}} \tag{3.7}$$

Vastaamattomuus aiheuttaa kuitenkin harhaa estimointiin. Bethlehemin mukaan ensimmäisen asteen Taylorin sarjakehitelmällä saadaan laskettua estimaattorin odotusarvo [Bethlehem (1988)]:

$$E(\overline{y}^*) \approx \frac{1}{N}\sum_{h=1}^{N} \frac{\rho_h Y_h}{\overline{\rho}} \tag{3.8}$$

Tällöin vastaamattomuusharha voidaan laskea estimaattorin ja oikean keskiarvon erotuksena. Yhdistämällä kaavat 3.7 ja 3.8 sekä lisäämällä kaksi toisensa kumoavaa tekijää $\left(\sum_{h=1}^{N} \rho_h \overline{Y} = \sum_{h=1}^{N} \overline{\rho}\overline{Y}\right)$ huomataan, että harha on estimoitavissa vastausalttiuden ja muuttujan välisellä kovarianssilla, kunhan se ensin jaetaan vastausalttiuksien keskiarvolla. [Bethlehem (1988)]:

$$\begin{aligned}
B(\overline{y}^*) &= E(\overline{y}^*) - \overline{Y} \\
&= \frac{1}{N\overline{\rho}}\sum_{h=1}^{N} \rho_h Y_h - \sum_{h=1}^{N} \overline{\rho} Y_h \\
&= \frac{1}{N\overline{\rho}}\sum_{h=1}^{N} \rho_h Y_h - \sum_{h=1}^{N} \overline{\rho} Y_h - \sum_{h=1}^{N} \rho_h \overline{Y} + \sum_{h=1}^{N} \overline{\rho}\overline{Y} \\
&= \frac{1}{N\overline{\rho}}\sum_{h=1}^{N} (Y_h - \overline{Y})(\rho_h - \overline{\rho}) \\
&= \frac{Cov(\rho, Y)}{\overline{\rho}}
\end{aligned} \tag{3.9}$$

Estimaattorin harhan yleinen määritelmä löytyy kappaleesta *3.2.6 Harhakorjattu R-indikaattori* (määritelmä 3.21).

3.1.4 Vastaamattomat

Puhuttaessa vastausharhasta ja edustavuudesta täytyy aina pitää mielessä se, että normaalissa tilanteessa ei ole minkäänlaisia takeita, että olettamukset vastaamattomista henkilöistä pitävät paikkansa. Mahdollisuus siihen, että puuttuneisuusmekanismi onkin MNAR, on kysymys, johon tulisi paneutua tarkemmin. Koska edustavuuslaskelmat perustuvat perusjoukkotason luokitteleviin apumuuttujiin, ei ole mitään takeita siitä, että valittuihin luokkiin kuuluvien vastaamattomien vastaukset vastaavat samoihin luokkiin kuuluneiden vastanneiden

vastauksia. Sinälläänhän tämä on ilmeistä, koska heidän vastauksiaan ei ole, niin ei niitä pääse mihinkään vertaamaan. Joissain tilanteissa toki päästään jälkikäteen näkemään, millainen oli todellisuus. Esimerkiksi vaalien jälkeen saadaan tieto siitä, ketkä valittiin sekä tieto siitä, ketkä äänestivät ja ketkä eivät äänestäneet. Tämän jälkeen äänestyskäyttäytymistä voidaan tarkastella perusjoukkotasolla. Voidaan toki kysyä, että mikä on kannatustutkimuksissa varsinainen kohdeperusjoukko? Sisältääkö kohdeperusjoukko koko kehikkoperusjoukon, eli kaikki äänioikeutetut vai vain ne, jotka haluavat äänioikeuttaan käyttää? Valinnasta riippumatta tiedossa on vain äänestäneiden mielipiteet, eli ”nukkuvien” ääni ei tässäkään kuulu. Joka tapauksessa saatujen tietojen perusteella voidaan säätää tulevien kannatusmittausten parametreja. Valitettavasti tällaista mahdollisuutta perusjoukon ”ruumiinavaukseen” ei kuitenkaan yleisesti ole tarjolla.

Joitain vuosikymmeniä sitten tämä vastaamattomuusongelma ei ollut vielä yhtä akuutti, sillä vastauskato oli tuolloin merkittävästi pienempää kuin nykyään [Plewes (2013)]. Vastauskadon kasvaessa on enenevässä määrin ryhdytty pohtimaan tätä vastaamattomien ja vastanneiden henkilöiden yhdenmukaisuutta. Suuri osa kadon korjausmenetelmistä (esim. imputointi, kalibrointi, balansointi) olettaa, että vastaamattomien ja vastanneiden jakaumat vastaavat toisiaan. Lisäpanostukset vaikeasti vastaavien ryhmän pienentämiseksi lienevät tulevaisuudessa entistäkin tarpeellisempia.

3.2 R-indikaattori

Edustavuus-indikaattorien tulisi olla tulkittavia, mitattavia ja normalisoitavia. R-indikaattori onkin yksi yritys tällaisen edustavuusmittarin luomisessa. RISQ-project, eli Representative Indicators for Survey Quality -project oli EU-rahoitteinen tutkimushanke, jonka yhtenä osana kehitettiin mm. R-indikaattori [RISQ 2007]. R-indikaattorin kehittäjät käyttivät alkuvaiheessa siitä välillä myös nimitystä R-indeksi: *R-indexes for the comparison of different fieldwork strategies and data collection modes.* R-indikaattoria kuvataan tässä pääluvussa sen kehittäjien Barry Schoutenin, Fannie Cobbenin and Jelke Bethlehemin kuvaamassa muodossa. [Schouten (2009b)]

3.2.1 Määritelmä

R-indikaattori (representativeness-indicator / edustavuusindikaattori) mittaa lopullisen havaintoaineiston edustavuuden poikkeamaa suhteessa teoreettiseen täyteen otokseen valittujen apumuuttujien suhteen. Eli R-indikaattori mittaa poikkeamaa täydellisestä heikosta edustavuudesta. Poikkeamamittana käytetään välille $[0, 1]$ skaalattua vastausalttiuden keskihajontaa, jossa tavallisen keskiarvon sijaan käytetään logistisella regressioanalyysillä estimoitua odotusarvoestimaattia. Täydellinen yhteensopivuus tuottaa arvon yksi ja vastaavasti täydellinen poikkeavuus arvon nolla. Eli R-indikaattorin arvo yksi tarkoittaa sitä, että kyselyyn vastanneiden jakauma vastaisi täysin otoksen jakaumaa valittujen apumuuttujien suhteen.

Määritelmä 3.10. *R-indikaattori*

$$R(\rho) = 1 - 2S(\rho)$$

R-indikaattorin laskemiseksi on siis ensin laskettava vaadittava etäisyys $S(\rho)$. Luonnollinen valinta etäisyysmitaksi on euklidinen etäisyys, joka käytännössä on vastausalttiuden ρ keskihajonta $S(\rho)$. Koska ρ on todennäköisyys, niin sen keskihajonnan maksimiarvo on puoli. Jotta R-indikaattorin vaihteluväliksi saadaan $[0, 1]$, vastausalttiuden keskihajonta täytyy kertoa kahdella ja saatu tulos tulee vielä vähentää yhdestä. Alun perin indikaattorista oli myös vaihtoehtoinen varianssipohjainen versio (R_2), jota tarkastellaan vaihtoehtoisten menetelmien yhteydessä (kaava 3.49).

3.2.2 Etäisyysmitan valinta

Geometriassa euklidinen etäisyys tarkoittaa ns. linnuntien käyttämistä, joka saadaan laskettua Pythagoraan lauseen avulla. Tilastotieteessä satunnaismuuttujan ja sen odotusarvon välinen poikkeama lasketaan euklidisena etäisyytenä ja tätä etäisyyttä kutsutaan perusjoukon keskihajonnaksi (σ). Perusjoukon keskihajonnan harhaton estimaattori on otoskeskihajonta (s), jonka vapausasteet putoavat yhdellä, koska sen laskemiseksi joudutaan otoksesta estimoimaan perusjoukon odotusarvo. Vastausindikaattorien tapauksessa keskihajonnan kaava muotoutuu seuraavanlaiseksi:

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \overline{\rho})^2} \tag{3.11}$$

R-indikaattorin tapauksessa kiinnostuksen kohteena on vastausalttiuksien ρ_i muokattu keskihajonta $S(\rho)$. Vastaajakohtainen vastausalttius ρ_i kuvaa vastaajan todennäköisyyttä vastata, sillä ehdolla että vastaaja on valikoitunut otokseen. [Schouten (2009b)]

Koska ρ on todennäköisyys, se vaihtelee välillä $[0, 1]$. Tästä seuraa, että keskihajonnan teoreettinen maksimiarvo on $\frac{1}{2}$, joka voidaan päätellä epäyhtälöstä:

$$S(\rho) \leq \sqrt{\overline{\rho}(1-\overline{\rho})} \leq \frac{1}{2}. \tag{3.12}$$

Teoreettisen maksiminsa $S(\rho)$ saavuttaa otoskoon kasvaessa kohti ääretöntä tilanteessa, jossa puolet ρ_i-arvoista on ykkösiä ja loput nollia. Tällaisessa tilanteessa R-indikaattori saa arvokseen 0. Tilanteessa, jossa kaikki vastaavat tai kukaan ei vastaa, keskihajonta on luonnollisestikin 0, kun minkäänlaista vaihtelua ei esiinny. Tästä seuraa, että R-indikaattori saa arvokseen 1. Vastausaste asettaa siis rajat R-indikaattorin mahdollisille arvoille. R-indikaattorin minimiarvo kasvaa, kun vastausaste etääntyy puolesta kumpaan tahansa suuntaan.

3.2.3 Yhteyden puuttuvuus -mitta

Kehittäjien mukaan R-indikaattoria voidaan pitää "*yhteyden puuttuvuus*" -mittana. Kun $R(\rho) = 1$, niin puuttuneisuuden ja tutkimuskohteen välillä ei ole minkäänlaista yhteyttä. R-indikaattorilla onkin yhteys Cramérin V-tunnuslukuun, jolla mitataan luokiteltujen muuttujien yhteensopivuutta. Palautetaan vielä mieleen, että Cramérin V vaihtelee välillä $[0, 1]$, siten että lähellä yhtä olevat arvot tarkoittavat voimakasta riippuvuutta.

Jos halutaan testata vastausalttiuden riippuvuutta apumuuttujien luokista, testaamiseen voidaan käyttää χ^2-yhteensopivuustestiä. Muokataan χ^2-yhtälöä hieman, jotta yhteys vastausalttiuden keskihajontaan saadaan hieman paremmin esiin. Kuvataan f_h:lla luokkaan h kuuluvien havaintojen suhteellista osuutta kaikista havainnoista, eli $f_h = N_h/N$. Oletetaan lisäksi, että kaikkien luokkaan h kuuluvien havaintojen vastausalttius on $\overline{\rho}_h$. Halutaan testata, onko vastausalttiuden ja luokittelevien muuttujien välillä riippuvuutta. Tällöin testisuure on:

$$\begin{aligned}
\chi^2 &= \sum_{h=1}^{H} \frac{\left(N_h\overline{\rho}_h - N_h\overline{\rho}\right)^2}{N_h\overline{\rho}} + \sum_{h=1}^{H} \frac{\left(N_h\left(1-\overline{\rho}_h\right) - N_h\left(1-\overline{\rho}\right)\right)^2}{N_h\left(1-\overline{\rho}\right)} \\
&= \sum_{h=1}^{H} \frac{N_h\left(\overline{\rho}_h - \overline{\rho}\right)^2}{\overline{\rho}} + \sum_{h=1}^{H} \frac{N_h\left(\overline{\rho}_h - \overline{\rho}\right)^2}{\left(1-\overline{\rho}\right)} \\
&= \sum_{h=1}^{H} \frac{Nf_h\left(\overline{\rho}_h - \overline{\rho}\right)^2}{\overline{\rho}\left(1-\overline{\rho}\right)} = \frac{N}{\overline{\rho}\left(1-\overline{\rho}\right)} \sum_{h=1}^{H} f_h\left(\overline{\rho}_h - \overline{\rho}\right)^2
\end{aligned} \tag{3.13}$$

Muokataan seuraavaksi vastausalttiuden varianssia $S^2(\tilde{\rho})$ samankaltaisempaan asuun.

$$S^2(\tilde{\rho}) = \frac{1}{N-1} \sum_{h=1}^{H} N_h\left(\overline{\rho}_h - \overline{\rho}\right)^2 = \frac{N}{N-1} \sum_{h=1}^{H} f_h\left(\overline{\rho}_h - \overline{\rho}\right)^2 \tag{3.14}$$

Yhdistämällä aiemmat yhtälöt havaitsemme, että χ^2-tunnusluku voidaan esittää myös muodossa:

$$\chi^2 = \frac{N}{\overline{\rho}\left(1-\overline{\rho}\right)} \frac{N-1}{N} S^2(\tilde{\rho}) = \frac{N-1}{\overline{\rho}(1-\overline{\rho})} S^2(\tilde{\rho}) \tag{3.15}$$

Cramérin V:tä laskettaessa χ^2-tunnusluku jaetaan otoskoolla, joka on kerrottu rivien ja sarakkeiden lukumääristä pienemmän vapausasteilla. Koska R-indikaattorin kyseessä ollessa toinen luokitus on dikotominen (vastaa/ei vastaa), niin jakajan kertoimeksi jää 1. Sijoitetaan vielä χ^2 arvo edellisestä yhtälöstä, niin saamme Cramérin V:n uudelleen muotoillun yhtälön.

$$V = \sqrt{\frac{\chi^2}{N\left(min(k,r)-1\right)}} = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\overline{\rho}(1-\overline{\rho})}} S(\tilde{\rho}) \tag{3.16}$$

Kun sijoitetaan yhtälö R-indikaattorin määritelmään, havaitaan, että R-indikaattori voidaan esittää Cramérin V:n avulla seuraavassa muodossa:

$$R(\rho) = 1 - 2V\sqrt{\frac{N\overline{\rho}(1-\overline{\rho})}{N-1}} \tag{3.17}$$

Kaavasta (3.17) havaitaan, että Cramérin V:n arvon noustessa R-indikaattorin arvo laskee. Tämä on tietenkin luonnollista, sillä mitä vahvempi yhteys luokkien ja vastaamisaktiivisuuden välillä on, sitä selkeämmin puuttuvat havainnot poikkeavat havaituista. [Schouten (2009b)]

3.2.4 Vastauspohjainen R-indikaattori

Edellisessä kappaleessa käsiteltiin teoreettista tilannetta, että tiedossa on koko otoksen vastausalttiudet. Todellisuudessa näin ei ole, vaan tiedetään vain, mikä on tilanne vastanneiden joukossa. Näillä tiedoilla saataisiin siis laskettua vain vastausalttiuden keskiarvo $\overline{\rho}$ ja vastaavasti vastausalttiuden varianssi $S^2(\rho)$ vastanneiden joukossa.

Jotta päästään käsiksi koko otokseen, täytyy havaittujen havaintojen sijaan käyttää jotain mallipohjaista estimaattia vastausalttiudelle, joka saadaan estimoitua myös vastaamattomille. Luonnollisia vaihtoehtoja ovat lineaarinen tai logistinen regressiomalli. Myös probit-regressiomallia voitaisiin käyttää tai jos esiintyy tarvetta, niin epäparametrisena vaihtoehtona voisi käyttää esimerkiksi CHAID-luokittelupuita. [Schouten (2009b)]

Jotta päästään estimoimaan puuttuvien havaintojen mahdollisia arvoja, käytetään mallintamisessa apuna apumuuttujia. Valittujen apumuuttujien arvot pitää tuntea koko otoksesta, eli sekä vastanneiden että vastaamattomien joukossa. Seuraavissa kaavoissa $\hat{\rho}_i$:t ovat estimoituja vastausalttiuden arvoja. Keskiarvoestimaatiksi saadaan tällöin:

$$\hat{\overline{\rho}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\rho}_i\frac{s_i}{\pi_i} \tag{3.18}$$

Estimoitu havaintokohtainen vastausalttius $\hat{\rho}_i$ kerrotaan sisältymisindikaattorilla s_i, joka jaetaan sisältymistodennäköisyydellä π_i. Sisältymistodennäköisyys on havainnon luokittelun mukainen todennäköisyys päätyä otokseen. Sisältymisindikaattori puolestaan saa arvokseen yksi, jos havainto kuuluu otokseen ja nollan, jos havainto ei kuulu otokseen. Käytännössä siis lasketaan yhteen vain otokseen kuuluvat havainnot, joiden painotettu summa jaetaan perusjoukon koolla. Vastauspohjaisen R-indikaattorin kaavaksi saadaan näillä merkinnöillä: [Schouten (2009b)]

Määritelmä 3.19. *Vastauspohjainen R-indikaattori*

$$R(\hat{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}\left(\hat{\rho}_i - \hat{\overline{\rho}}\right)^2}$$

3.2.5 Logistinen regressiomalli

Koska on vain kaksi mahdollista vaihtoehtoa, eli henkilö joko vastaa tai ei vastaa, on vastemuuttuja dikotominen. Dikotomisen vasteen kyseessä ollessa ilmeinen valinta mallintamiseen on logistinen regressioanalyysi. Logistisen regressiomallin oletukset ovat:

1. Selitettävän muuttuja tulee olla dikotominen tai järjestysasteikollinen
2. Havaintojen pitää olla riippumattomia
3. Selittäjät eivät saa olla vahvasti korreloituneita keskenään
4. Vedonlyöntikertoimen logaritmin (log odds) ja selittäjien välillä tulee olla lineaarinen yhteys

Jonkinlaisen nyrkkisääntönä pidetään myös sitä, että havaintoja tulisi olla vähintään kymmenen kertaa selittävien muuttujien määrä. Suurilla otoskoilla dikotomiset selittäjät ovat myös sallittuja, koska niillä on kvalitatiivisuudestaan huolimatta kvantitatiivisesti määrittyvä ”ollako vai eikö olla”-ominaisuus, eli kyseinen asia joko on olemassa tai ei ole olemassa. Tämä ominaisuus mahdollistaa järjestys- ja laatueroasteikollisten muuttujien käyttämisen selittäjinä logistisessa regressiomallissa.

Järjestys- ja laatueroasteikolliset selittäjät täytyy siis korvata dikotomisilla indikaattorimuuttujilla, jotta niitä voidaan käyttää selittäjinä. Eli kaikille muuttujan arvoille luodaan oma muuttuja, joka saa arvon 1, kun alkuperäisellä muuttujalla on haluttu arvo ja muulloin arvon 0. Näitä muuttujia tarvitaan yksi vähemmän, kun alkuperäisellä muuttujalla on luokkia, sillä viimeiseen luokkaan kuuluvuus saadaan kuvattua edellisten muuttujien lineaarikombinaatiolla. Jos kyseessä on järjestysasteikollinen muuttuja, niin tässä operaatiossa valitettavasti menetetään muuttujan sisältämä järjestysinformaatio.

Määritelmä 3.20. *Logistinen regressiomalli*

$$\rho = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}+1}$$

Logistisella regressiomallilla estimoidaan siis vastausalttiudelle ρ odotusarvo valittujen apumuuttujien suhteen. Pyritään siis löytämään aineistosta sopivia apumuuttujia, jotka jakavat aineiston toisensa poissulkeviin osajoukkoihin. Kyseeseen tulevat vain sellaiset muuttujat, joiden arvot ovat saatavissa koko aineistolle, eli sekä varsinaisiin kysymyksiin vastanneille että vastaamattomille.

Valitut apumuuttujat asetetaan logistisen regressiomallin selittäjiksi (x_i). Selitettävänä muuttujana on vastausalttius ρ. Tässä mallissa arvolla 1 kuvataan varsinaiseen kyselyyn vastaamista ja arvolla 0 kyselyyn vastaamattomuutta. Mallin kertoimet (β_i) estimoidaan suurimman uskottavuuden menetelmällä. Lopulliseen malliin jätetään selittäjistä ne, jotka ovat merkitseviä. Logistisessa regressiomallissa selittäjien merkitsevyyttä voidaan testata mm. Waldin testillä tai vaihtoehtoisesti tarkastelemalla vetosuhteiden (odds ratio) luottamusvälejä. Jos arvo yksi mahtuu

kyseisen selittäjän vetosuhteen 95 %:n luottamusvälille, niin selittäjä ei ole merkitsevä. Valitettavasti logistiselle regressiomallille ei löydy selitysasteen kaltaista selkeästi tulkittavaa hyvyysmittaria. Erilaisten mallien hyvyyttä voidaan kuitenkin vertailla mm. uskottavuusosamäärätestillä.[Hosmer(2013)]

3.2.6 Harhakorjattu R-indikaattori

Luodessaan hypoteesien testauksen artikkelissaan *"Contributions to theory of testing statistical hypotheses"* vuonna 1936 Jerzy Neyman ja Egon Pearson loivat käsitteen harhattomuus. Kaksi vuotta myöhemmin Jerzy Neyman ja Florence Nightingale David määrittelivät artikkelissaan *"Extension of the Markoff theorem on least squares"* piste-estimaattien harhattomuuden, siten että estimaatti on harhaton, jos sen odotusarvo on yhtä suuri kuin sen estimoitu arvo [Lehmann (1951)]. Harhattomuuden olemassa olosta voitaneen päätellä, että on olemassa myös harha. Yleisesti estimaattorin harha $B(T)$ saadaan määritettyä, kun parametrin oikea arvo θ vähennetään siitä lasketun estimaattorin odotusarvosta $E(T)$.

Määritelmä 3.21. *Harha*

$$B(T) = E(T) - \theta$$

Vastausalttiuksien ρ_i estimoinnista johtuen R-indikaattori, ei ole harhaton estimaattori, vaan siinä esiintyy otoskoosta riippuvaa harhaa, siten että otoskoon pienentyessä harha kasvaa [Shlomo (2009a)]. Yksinkertaista satunnaisotantaa käytettäessä harhakorjattu R-indikaattori lasketaan kaavalla:

$$\hat{R}(\hat{\rho}_X) = 1 - 2\sqrt{\left(1 + \frac{1}{n} - \frac{1}{N}\right)\hat{S}^2\left(\hat{\rho}_X\right) - \frac{N}{n}\sum_{i\in S}\nabla h\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^2\boldsymbol{x}_i\left[\sum_{j\in S}\nabla h\left(\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_j\boldsymbol{x}_j'\right]^{-1}\boldsymbol{x}_i}, \tag{3.22}$$

Kaavassa ∇h on gradienttivektori, joka koostuu ensimmäisen kertaluvun derivaatoista $\hat{\boldsymbol{\beta}}$:n suhteen. Koska $\hat{\boldsymbol{\beta}}$-vektorin estimointiin käytetään logistista regressiomallia, niin linkkifunktion h gradientti määrittyy kaavalla:

$$\nabla h(\boldsymbol{x}'\hat{\boldsymbol{\beta}}) = \frac{e^{\boldsymbol{x}'\hat{\boldsymbol{\beta}}}}{\left(1 + e^{\boldsymbol{x}'\hat{\boldsymbol{\beta}}}\right)^2} \tag{3.23}$$

Harhakorjatun R-indikaattorin laskentakaava, kun käytetään osittaista otantaa, on: [Shlomo (2009b)]

$$\hat{R} = 1 - 2\sqrt{\hat{S}^2 + \sum_{h=1}^{H}\frac{N_h^2}{N^2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\hat{S}_h^2 - \frac{\sum_{i\in}S\frac{N_{h(i)}}{n_{h(i)}}\nabla h\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^2\boldsymbol{x}_i\left[\sum_{j\in S}\nabla h\left(\boldsymbol{x}_j'\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_j\boldsymbol{x}_j'\right]^{-1}\boldsymbol{x}_i}{N}} \tag{3.24}$$

Kaavassa: $\hat{S}_h^2(\hat{\rho}_X) = \frac{1}{n_h - 1} \sum_{s_h} \left(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,h}\right)^2$, $\hat{\bar{\rho}}_{X,h} = \frac{1}{n_h} \sum_{s_h} \hat{\rho}_X(x_i)$

Shlomo ja Schouten osoittivat massiivisilla simulaatiokokeilla, että alle 15000 havainnon otoksille kannattaa käyttää harhakorjattua R-indikaattoria. Tätä suuremmille otoksille voidaan käyttää tavallista R-indikaattoria. [Shlomo (2013)]

3.2.7 R-indikaattorin hajonnan estimointi

RISQ-projektin ensimmäisessä R-indikaattorin laskentakoodissa hajonnan estimointiin käytettiin ei-parametrista bootstrap-menetelmää, eli palauttaen tehtyä toistettua uudelleenotantaa otoksesta. Oletetaan, että alkuperäisessä otoksessa on B kappaletta havaintoja. Poimitaan alkuperäisestä otoksesta B kappaletta havaintoja palauttaen ja lasketaan saadulle uudelle otokselle keskihajonta. Toistetaan uudelleen poimintaa ja keskihajonnan laskentaa riittävän kauan, jotta keskihajontojen keskiarvo vakiintuu. Bootstrap-menetelmällä laskettu R-indikaattorin keskihajonta on: [Schouten (2009b)]

$$S(R)^{BT} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{R}_b^{BT} - \hat{\overline{R}}^{BT}\right)^2} \tag{3.25}$$

Kaavassa $\hat{\overline{R}}^{BT}$ on keskimääräinen otoksesta estimoitu R-indikaattori:

$$\hat{\overline{R}}^{BT} = \frac{1}{B} \sum_{b=1}^{B} \hat{R}_b^{BT} \tag{3.26}$$

Vaihtoehtoinen tapa hajonnan määrittämiseksi on käyttää jackknife-menetelmää, jossa otoksesta jätettään aina yksi havainto pois ja lasketaan R-indikaattori tälle uudelle aineistolle. Tämä laskenta toistetaan siten, että jokainen havainto on tullut kerran pois jätetyksi. Näin saadun R-indikaattoriaineiston avulla lasketaan sitten keskihajontaestimaatti alkuperäiselle R-indikaattorille. Kaavassa (3.27) n on sekä alkuperäisten havaintojen että uusien jackknife-otosten lukumäärä.

$$S(R)^{JK} = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \left(R_i^{JK} - \overline{R}^{JK}\right)^2} \tag{3.27}$$

R-indikaattorille on johdettu myös analyyttinen varianssiestimaattori, joka on määritelty Shlomon, Skinnern ja Schoutenin kirjoittamassa artikkelissa: *Estimation of an indicator of the Representativeness of Survey Response* lehdessä *Journal of Statistical Planning and Inference*. Tätä analyyttistä varianssiestimaattoria käytetään RISQ-projektin lopullisessa R-indikaattorin laskentakoodissa. [Shlomo (2013)]

Riippumatta siitä kuinka hajonta estimoidaan, niin R-indikaattorin luottamusvälin laskentaan käytetään seuraavaa kaavaa:

$$1 - 2\sqrt{\tilde{S}_\rho^2 \pm z_{1-\alpha/2} \sqrt{\nu(\hat{S}_\rho^2)}} \tag{3.28}$$

3.2.8 Tulkinta

R-indikaattorin selkein käyttötarkoitus on aineiston edustavuuden kehityksen seuranta perättäisillä ajanhetkillä tai jo aineiston hankinnan aikana. Itse suoran R-indikaattoriaikasarjan lisäksi kannattaa toki tutkailla myös muutoksia suhteessa indikaattorin luottamusväliin. Jos taas lähdetään hakemaan tulkintaa yksittäiselle R-indikaattorille, niin täytyy ensin määrittää yhteys vastaamattomuusharhan ja edustavuuden välille. Mielivaltaisesti valitun tutkimuskohteen keskiarvon $\hat{\bar{y}}_r$ standardoidun harhan maksimi määrittyy kaavalla:

$$\frac{|B(\hat{\bar{y}}_r)|}{S(y)} = \frac{|Cov(y, \rho_y)|}{S(y)\rho} = \frac{|Cov(y, \rho_\aleph)|}{S(y)\rho} \leq \frac{S(\rho_\aleph)}{\rho} = \frac{1 - R(\aleph)}{2\rho} \tag{3.29}$$

Yllä nähtävä harhan B suhde kovarianssin ja vastausalttiuden osamäärään on määritelty kaavassa 3.9 jo aiemmin. $\aleph$ viittaa supervektoriin, joka koostuu apumuuttujista, joiden avulla vastaajan vastauskäyttäytyminen selviäisi täydellisesti. $\rho_\aleph$:ia ei luonnollisestikaan tunneta, mutta sitä voidaan estimoida käytössämme olevalla ρ_X:llä. Näillä oletuksilla maksimaalista absoluuttista standardoitua harhaa B_m voidaan estimoida kaavalla:

$$B_m(X) = \frac{1 - R(X)}{2\rho} \tag{3.30}$$

Toinen mielenkiinnon kohde, jota voidaan tutkia yksittäisen R-indikaattorin avulla, on maksimaalinen kontrasti vastaajien ja vastaamattomien välillä. Jotta saadaan harha selville, lasketaan kyselyyn vastanneiden ja kyselyyn vastaamattomien antamien vastausten odotusarvojen erotus kiinnostuksen kohteena olevasta muuttujasta ja kerrotaan tämä vastaamattomuusasteella:

$$B(\hat{\bar{y}}_r) = (1 - \rho)(E(\hat{\bar{y}}_r) - E(\hat{\bar{y}}_{nr})) \tag{3.31}$$

On tietenkin selvää, että vastaamattomien vastaukset eivät käytännössä ole saatavissa, joten kyse on tämän teoreettisen suureen estimoinnista. Yhdistämällä kaavat 3.29 ja 3.31 saamme laskettua estimaatin maksimaaliselle absoluuttiselle standardoidulle kontrastille, kun maksimaalinen absoluuttinen standardoitu harha jaetaan vastaamattomuusasteella:

$$C_m(X) = \frac{1 - R(X)}{2\rho(1 - \rho)} \tag{3.32}$$

R-indikaattori, maksimaalinen harha ja maksimaalinen kontrasti tarjoavat meille mahdollisuuden tutkia kerätyn aineiston laatua edustavuuden kannalta. Määritellään edellisten pohjalta kolme vastausedustavuusfunktiota: RR_1, RR_2 ja RR_3 (response-representativity functions). Näiden avulla pystytään määrittelemään R-indikaattorille jonkinlaiset hyväksyttävyyden alarajat. Ensimmäinen vastausedustavuusfunktio rajaa maksimaalista hajontaa vastausalttiudessa:

$$RR_1(\gamma, \rho) = 1 - \frac{2}{\zeta_{1-0.05\alpha}}\gamma \tag{3.33}$$

Toinen vastausedustavuusfunktio rajaa maksimaalista harhaa:

$$RR_2(\gamma, \rho) = 1 - 2\rho\gamma \tag{3.34}$$

Kolmas vastausedustavuusfunktio rajaa maksimaalista kontrastia:

$$RR_3(\gamma, \rho) = 1 - 2\rho(1-\rho)\gamma \tag{3.35}$$

Kaavassa RR_1 $\zeta_{1-0.05\alpha}$ on 95 %:n luottamusvälin laskentaan tarvittava arvo normaalijakauman kertymäfunktiosta. Kaavassa RR_1 γ on valittu kynnysarvo, joka määrittää maksimaalisen poikkeaman vastausalttiudesta. Vastaavasti kaavassa RR_2 γ on kynnysarvo harhalle ja kaavassa RR_3 kontrastille. RISQ-projektin käyttämissä esimerkeissä γ:n arvoiksi on valittu 5, 10 ja 20 prosenttia. RR_1 on aineistosta riippumaton mittari, jonka arvo annetussa muodossa, eli 95 %:n luottamusväliä käytettäessä, on hyvin lähellä $(1-\gamma)$:a. [Schouten (2009a)]

3.2.9 Osittaiset R-indikaattorit

Osittaiset R-indikaattorit määritellään yksittäisen apumuuttujan tai sen mahdollisten luokkien suhteen. Niiden tarkoitus on mitata, paljonko yhdellä muuttujalla tai muuttujan luokalla on vaikutusta vastausalttiuksien vaihteluun. Eräs spesifi tarkoitus on selvittää, mitkä luokat ovat yli- tai aliedustettuina toteutuneessa havaintoaineistossamme.

Vastausalttiuden varianssi voidaan jakaa kahteen osaan muuttujan sisäiseen ja muuttujien väliseen varianssiin. Kaavoissa W on luokitteleva muuttuja, joka jakaa aineistoon L:n luokkaan ja U vastaavasti on perusjoukon havaintojen lukumäärä kyseisessä luokassa. Muuttujan sisäinen varianssi saadaan laskettua kaavalla:

$$S_w^2\left(\rho_X|W\right) = \frac{1}{N-1}\sum_{l=i}^{L}\sum_{i\in U_l}\left(\rho_X(x_i) - \overline{\rho}_{X,l}\right)^2 \tag{3.36}$$

Kun halutaan laskea keskimääräinen sisäinen varianssi yhdessä W:n määrittämässä luokassa, niin se saadaan laskettua jättämällä pois ulompi summaus:

$$S_w^2\left(\rho_X|W=l\right) = \frac{1}{N-1}\sum_{i\in U_l}\left(\rho_X(x_i) - \overline{\rho}_{X,l}\right)^2 \tag{3.37}$$

Yllä mainittujen varianssien estimaatit saadaan laskettua seuraavilla kaavoilla, joissa S_l on otoksen havaintojen lukumäärä kyseisessä luokassa ja d_i on kyseisen luokan painokerroin:

$$\hat{S}_w^2\left(\hat{\rho}_X|W\right) = \frac{1}{N-1}\sum_{l=i}^{L}\sum_{i\in S_l} d_i\left(\hat{\rho}_X(x_i) - \hat{\overline{\rho}}_{X,l}\right)^2 \tag{3.38}$$

$$\hat{S}_w^2\left(\hat{\rho}_X|W=l\right) = \frac{1}{N-1}\sum_{i\in S_l} d_i\left(\hat{\rho}_X(x_i) - \hat{\overline{\rho}}_{X,l}\right)^2 \tag{3.39}$$

Vastaavin merkinnöin muuttujien välinen varianssi saadaan laskettua kaavalla:

$$S_b^2\left(\rho_X|W\right) = \frac{1}{N-1}\sum_{l=1}^{L} N_l\left(\overline{\rho}_{X,l} - \overline{\rho}_X\right)^2 \cong \sum_{l=1}^{L}\frac{N_1}{N}\left(\overline{\rho}_{X,l} - \overline{\rho}_X\right)^2 \tag{3.40}$$

Ja edelleen yhden luokan keskimääräinen osuus luokkien välisestä varianssista saadaan laskettua kaavalla:

$$S_b^2\left(\rho_X|W=l\right) = \frac{N_1}{N}\left(\overline{\rho}_{X,l} - \overline{\rho}_X\right)^2 \tag{3.41}$$

Ja näiden estimaatit saadaan laskettua seuraavilla kaavoilla:

$$S_b^2\left(\rho_X|W\right) = \sum_{l=1}^{L}\frac{\hat{N}_l}{N}\left(\hat{\overline{\rho}}_{X,l} - \hat{\overline{\rho}}_X\right)^2 \tag{3.42}$$

$$S_b^2\left(\rho_X|W\right) = \frac{\hat{N}_l}{N}\left(\hat{\overline{\rho}}_{X,l} - \hat{\overline{\rho}}_X\right)^2 \tag{3.43}$$

Edellä määriteltyjen varianssiestimaattorien avulla saadaan johdettua ehdollistamaton osittainen R-indikaattori tilanteessa, jossa muuttujaa Z ei käytetä vastausalttiuden estimointiin:

$$P_u(Z,k,\rho_X) = S_b\left(\overline{\rho}_X|Z=k\right)\frac{\left(\overline{\rho}_{X,k} - \overline{\rho}_X\right)}{\left|\overline{\rho}_{X,k} - \overline{\rho}_X\right|} = \sqrt{\frac{N_k}{N}}\left(\overline{\rho}_{X,k} - \overline{\rho}_X\right) \tag{3.44}$$

Kun muuttujaa Z käytetään vastausalttiuden estimointiin, niin kaava muuntuu muotoon:

$$P_u(Z,k,\rho_{X,Z}) = S_b\left(\overline{\rho}_{X,Z}|Z=k\right)\frac{\left(\overline{\rho}_{X,Z,k} - \overline{\rho}_X\right)}{\left|\overline{\rho}_{X,Z,k} - \overline{\rho}_X\right|} = \sqrt{\frac{N_k}{N}}\left(\overline{\rho}_{X,Z,k} - \overline{\rho}_{X,Z}\right) \tag{3.45}$$

Ehdollistamattomia osittaisia R-indikaattoreita edellä mainituissa tilanteissa estimoidaan seuraavilla kaavoilla:

$$\hat{S}_b\left(\hat{\rho}_X|Z=k\right) = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\overline{\rho}}_{X,k} - \hat{\rho}_X\right) \tag{3.46}$$

$$\hat{S}_b\left(\hat{\rho}_{X,Z}|Z=k)\right) = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\overline{\rho}}_{X,Z,k} - \hat{\rho}_{X,Z}\right) \tag{3.47}$$

Näiden osittaisten ehdollistamattomien R-indikaattorien vaihteluväli on $P_u \in [-1,1]$ toisin kuin tavallisen R-indikaattorin. Tästä ehdollistamattomasta osittaisesta R-indikaattorista on olemassa myös välillä $[0,1]$ vaihteleva versio, joka saadaan yksinkertaisesti ottamalla itseisarvo tässä esitellystä indikaattorista. Tällä jälkimmäisellä ehdollistamattomalla osittaisella R-indikaattorilla saadaan laskettua myös muuttujatasoisia tuloksia.

Lopuksi esitellään vielä ehdollistettu osittainen R-indikaattori. Tässä tilanteessa muuttujan Z on oltava yksi luokittelevista muuttujista. Tällöin ositus tehdään X muuttujan mukaan, kun aiemmin se aina tehtiin Z muuttujan mukaan.

$$P_c(Z,k,\rho_{X,Z}) = \sqrt{\frac{1}{N-1}\sum_{l=1}^{L}\sum_{i\in U_l}\delta_{k,i}\left(\rho_{X,Z}\left(x_i,z_i\right) - \overline{\rho}_{x,z,l}\right)^2} \tag{3.48}$$

Koska R-indikaattorissa esiintyy otoskoosta riippuvaista harhaa, niin osittaisissa R-indikaattoreissakin esiintyy otoskoosta riippuvaista harhaa, joka kasvaa otoskoon pienentyessä. RISQ-projektissa osittaisten R-indikaattorien harhakorjaus perustuu pro rating -metodiin, jossa R-indikaattorin harhan ajatellaan jakautuvan varianssihajotelman eri osioille. Harhan korjausta ei suositella osittaisissa R-indikaattoreissa, kun otoskoko on yli 15000. Pienemmillä otoksilla kannattaa käyttää harhakorjattuja R-indikaattoreita. [Shlomo (2012)]

3.2.10 Vaihtoehtoja

Erilaisia vaihtoehtoisia tunnuslukuja puuttuneisuuden vaikutuksen mittaamiseen on useita. Tähän mennessä esillä ovat olleet, R-indikaattori sekä sen laskennassa tarvittava vastausalttius ja luonnollisesti myös vastausaste. Muita puuttuneisuuden vaikutuksen mittareita ovat muun muassa:

- Varianssiin perustuva R-indikaattori R_2
- Uskottavuusfunktioon perustuva R-indikaattori R_3
- Vastausalttiuksien variaatiokerroin CV
- Apumuuttujien ja puuttuneisuuden riippuvuusmitta V
- Absoluuttisen harhan mediaani MAB
- BIX-indikaattori BIX
- Harhaindikaattori Q^2
- Epätasapainoestimaattori IMB

Näistä mittareista neljä ensimmäistä perustuu R-indikaattorin tavoin vastausalttiuden estimointiin logistisella regressioanalyysillä. Muut mittarit perustuvat havaintolukumääriin apumuuttujien määräämissä luokissa.

Varianssiin perustuva R_2-indikaattori (kaava 3.49) on laskennallisesti yhteneväinen R-indikaattorin kanssa muilta osin, paitsi että se perustuu varianssiin keskihajonnan sijasta. R_2-indikaattorin suhde R-indikaattoriin onkin suunnilleen sama kuin keskihajonnan suhde varianssiin. Toki sillä poikkeamalla, että myös R_2-indikaattori skaalataan välille $[0, 1]$. Poikkeavien luokkien vaikutus on siis suurempi ja saadut lukuarvot ovat muutenkin äärimmäisempiä. Eli puhtaasti laskennallisista syistä R_2-indikaattori antaa lähempänä nollaa tai yhtä olevia arvoja verrattuna R-indikaattoriin. [Schouten (2007)]

$$R_2(\rho) = 1 - 4S^2(\rho) \tag{3.49}$$

R_3-indikaattori (kaava 3.51) poikkeaa aiemmista siinä, että se ei perustu vastausalttiuden varianssiin. R_3-indikaattori perustuu PRE-testiin (Proportional Reduction of Error), jolla voidaan mitata lineaarisessa regressiossa mallin selittämää vaihtelua. Tarkemmin ottaen R_3-indikaattori perustuu Nagelkerken pseudo-$\Re^2$ -tunnuslukuun (kaava 3.50), joka on suunniteltu luokitteleville selittäville muuttujille. Schoutenin ja Cobbenin paperissa *"R-indexes for the comparison of different fieldwork strategies and data collection modes"*, jossa määritellään R_3-indikaattori, Nagelkerken pseudo-$\Re^2$ -tunnusluvun kaavassa on ylimääräinen toinen potenssi, kun sitä verrataan muihin löydettyihin lähteisiin. Tämä potenssiin korotus tekee sen, mitä potenssiin korotus tekee, nollan ja yhden välillä vaihteleville muuttujille, eli indikaattorin arvoista tulee äärimmäisempiä. Kaavassa 3.50 on Nagelkerken pseudo-$\Re^2$ -tunnusluvusta käytetty Nagelkerken alkuperäistä muotoilua ja myöhemmät laskennat perustuvat myös tähän kaavaan. [Nagelkerke (1991)]

$$\Re^2 = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - (L_0)^{2/n}} \tag{3.50}$$

Kaavassa $L_0 = \hat{\bar{\rho}}^{n\hat{\bar{\rho}}}(1-\hat{\bar{\rho}})^{n(1-\hat{\bar{\rho}})}$ ja $L_1 = \prod_{h=1}^{H} \hat{\bar{\rho}}_h^{n_h\hat{\bar{\rho}}_h}(1-\hat{\bar{\rho}}_h)^{n_h(1-\hat{\bar{\rho}}_h)}$. Indikaattorin suora laskenta ei onnistu, sillä sekä L_0 että L_1 menevät esimerkiksi SAS-ohjelman tarjoamalla laskutarkkuudella nolliksi. Hieman eksponentteja pyörittelemällä ja logaritmoimalla tulokset saadaan kuitenkin laskettua. Lopullinen indikaattori (kaava 3.51) saadaan, kun Nagelkerken pseudo-$\Re^2$ vähennetään yhdestä. Indikaattorin arvo vaihtelee välillä $[0, 1]$. [Schouten (2007)]

$$\hat{R}_3(\tilde{\rho}) = 1 - \Re^2 \tag{3.51}$$

Vastausalttiuksien variaatiokerroin CV (Coefficient of Variation) (kaava 3.52) on mittayksikötön tunnusluku, joka kuvaa logistisella regressioanalyysillä estimoidun vastausalttiuden vaihtelua. Vastausalttiuden variaatiokerroin vaihtelee välillä $[0, 1]$. Mitä pienempi arvo on, niin sitä paremmin saadut vastaukset muistuttavat

perusjoukkoa. Vastausalttiuden variaatiokerroin on yhtenevä maksimaalisen absoluuttisen harhan estimointiin käytetyn B_m:n kanssa, joka määriteltiin kaavassa 3.30.

$$CV = \frac{S(\rho)}{\overline{\rho}} \tag{3.52}$$

Aiemmin mainittiin, että R-indikaattorilla ja laatueroasteikollisten muuttujien riippuvuustunnusluku Cramérin V:llä on funktionaalinen yhteys. Tästä seuraa, että apumuuttujien muodostamien luokkien ja vastausalttiuden välistä riippuvuusmittaa voidaan myös pitää jonkinlaisena mittarina sille, että onko puuttuneisuus liian suurta. Cramérin V:n (kaava 3.53) laskenta perustuu χ^2-tunnuslukuun ja se vaihtelee välillä $[0, 1]$. Toisin kuin tavallisessa χ^2-tunnusluvun laskennassa tässä odotetut arvot on estimoitu logistisella regressiomallilla. Lähellä nollaa olevat arvot kertovat siitä, että puuttuneisuus on samanlaista kaikissa apumuuttujien muodostamissa luokissa. Cramérin V:n laskennallinen yhteys R-indikaattorin esiteltiin jo aiemmin luvussa 3.2.3.

$$V = \sqrt{\frac{\sum_{h=1}^{H} n_h \left(\rho_h - \overline{\rho}\right)^2}{n\overline{\rho}(1-\overline{\rho})}} \tag{3.53}$$

MAB (Median Absolut Bias) (kaava 3.54) on osallistumattomuusharhan itseisarvojen mediaani. Lasketaan siis kaikista luokista harha ja etsitään saaduista luvuista se, jonka itseisarvo on suuruusjärjestyksessä keskimmäinen. Koska mitataan aineiston harhaisuutta, niin pieni arvo kuvaa hyvää tilannetta. Ja koska kyse on suhteellisten osuuksien erotuksien itseisarvoista, niin arvo vaihtelee välillä $[0, 1]$. [Cornesse (2018)]

$$MAB = \underset{1 \leq h \leq H}{med} \left| \left(\frac{n_{rh}}{n_r}\right) - \left(\frac{n_h}{n}\right) \right| \tag{3.54}$$

BIX-indikaattorin (kaava 3.55) tavoitteena on olla yksinkertaisesti laskettava ja helposti tulkittava mitta harhalle. Laskennan kannalta se toki sitä on, koska kyse on havaittujen ja odotettujen arvojen erotusten itseisarvojen summasta apumuuttujien määräämissä luokissa. Valitettavasti indikaattorin määrittelevä Zhangin artikkeli ei kuitenkaan anna saadulle indikaattorille mitään suoraa tulkintaa, vaan sitä käytetään eri ajanhetkien tulosten vertailuun kuten R-indikaattoria. BIX on ”lyhenne” (Bias Indicator conditioning on the combination of the X:s), eli harhaindikaattori x-muuttujien määrittelemissä luokissa. BIX-indikaattorin arvot vaihtelevat välillä $[0, 1]$. [Zhang (2013)]

$$BIX = \sum_{h=1}^{H} \left| \left(\frac{n_{rh}}{n_r}\right) - \left(\frac{n_h}{n}\right) \right| \tag{3.55}$$

Särndalin ja Lundströmin Q^2 (kaava 3.56) ei varsinaisesti ole edustavuusindikaattori, mutta sitä voidaan vertailumielessä pitää sellaisena. Q^2:n alkuperäisenä käyttötarkoituksena on löytää mahdollisimman hyvä apumuuttujajoukko vastaamattomuusharhan minimoimiseksi. [Särndal (2005)]

$$Q^2 = \sum_{h=1}^{H} \frac{n_h^2}{n_r \cdot n_{rh}} - \frac{n^2}{n_r^2} \tag{3.56}$$

Epätasapainoestimaattori IMB (Imbalance estimator) (kaava 3.57) kuvaa sitä, kuinka hyvin otoksen keskiarvo vastaa perusjoukon vastaavaa tunnuslukua. IMB-estimaattorin vaihteluväli on $[0, p(1-p)]$, eli estimaattorin arvo on maksimissaan $\frac{1}{4}$. Estimaattori on tarkoitettu aineiston tasapainottamiseen aineistonkeruuvaiheessa, siten että vastauskadosta kärsiviä luokkia täydennetään, kunnes tasapaino on riittävällä tasolla. Tässä yhteydessä pääkäyttötarkoituksestaan poiketen estimaattoria käytetään edustavuusindikaattorina puhtaasti vertailumielessä. [Särndal (2016)]

$$IMB = \sum_{h=1}^{H} \frac{n_h}{n} \left(\frac{n_{rh}}{n_h} - \frac{n_r}{n} \right)^2 \tag{3.57}$$

Luku 4

Tulokset

4.1 Kadon ja edustavuuden kehitys

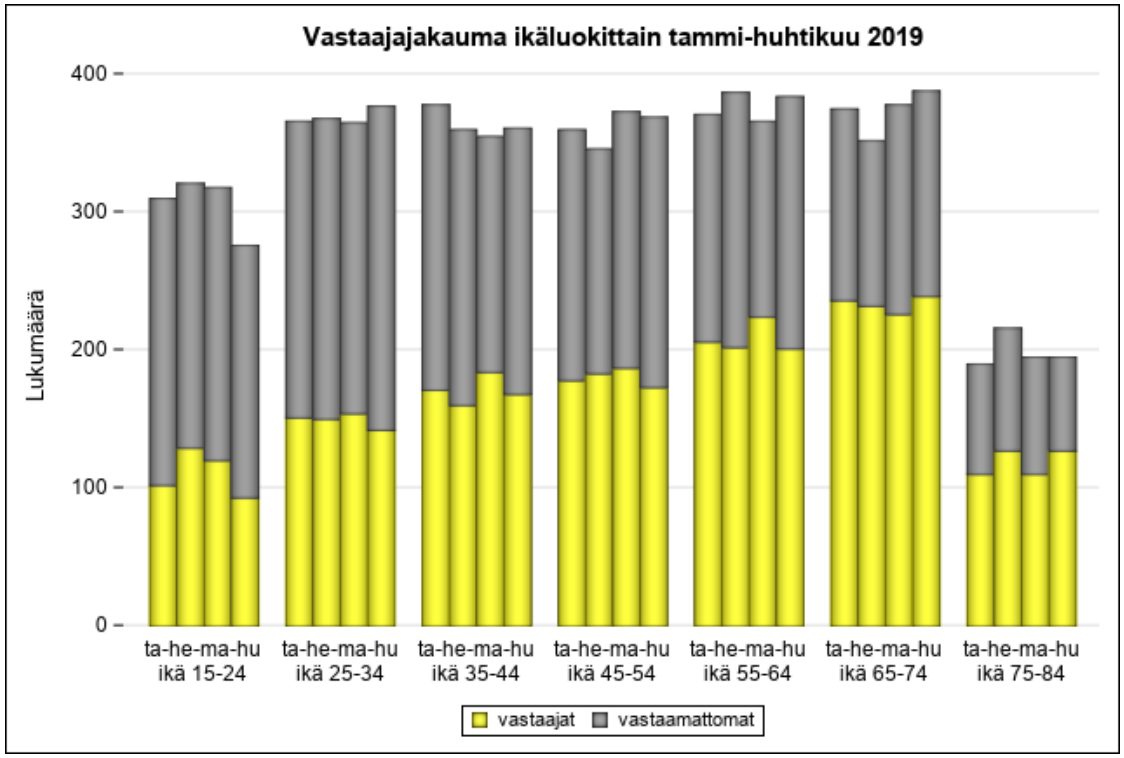
Kadon patoaminen on kyselytutkimusten tekijöille ns. tuhannen taalan kysymys. Kuluttajabarometrin tapauksessa tähän pyritään siirtymällä puhelinhaastattelusta yhdistelmätiedonkeruuseen. Pelkkä vastaajien lisääminen ei kuitenkaan riitä, vaan vastaajajoukon tulee myös edustaa perusjoukkoa mahdollisimman hyvin. Tämän takia on tarpeen tutkia edustavuuden kehitystä. Tässä tutkimuksessa edustavuutta tarkastellaan R-indikaattorin avulla. R-indikaattorin lisäksi tarkastellaan myös muita vaihtoehtoisia edustavuusmittareita. Näitä mittareita vertaillaan pääasiassa R-indikaattoriin ja vastausalttiuteen, mutta toki niiden keskinäisiäkin riippuvuuksia hieman tarkastellaan.

Täydellistä edustavuutta on vaikea saavuttaa tai edes testata, joten tässä tutkitaan edustavuutta suhteessa valittuihin apumuuttujiin. Näiden apumuuttujien arvot pitää tietää sekä kyselyyn vastanneiden että vastaamattomien joukossa. Kuluttajabarometriaineistossa on helposti saatavilla neljä tällaista muuttujaa, joita tarkastellaan lähemmin seuraavassa luvussa.

4.1.1 Apumuuttujat

Vanhassa kuluttajabarometriaineistossa on neljä apumuuttujaa, joihin löytyy tiedot kaikilta otokseen osuneilta. Näiden apumuuttujien arvot pitää tuntea sekä kyselyyn vastanneilta että vastaamattomilta, jotta niiden avulla voidaan myöhemmin estimoida vastausalttiutta. Nämä apumuuttujat ovat: sukupuoli, asuinalue, koulutustaso ja ikä. Näistä apumuuttujista ainoastaan ikä ja sukupuoli ovat mitta-asteikkojensa puolesta suoraan käytettävissä selittävinä muuttujina logistisessa regressiomallissa. Järjestysasteikollinen koulutusmuuttuja ja laatueroasteikollinen asuinalue tulee uudelleen luokitella sarjaksi toisensa poissulkevia dikotomisia indikaattorimuuttujia, ennen kuin niitä voi käyttää logistisessa regressiomallissa selittäjinä. Tässä uudelleen luokittelussa valitettavasti menetetään hieman informaatiota, kun järjestysasteikollinen koulutusmuuttuja menettää järjestysominaisuutensa.

Sekä taulukossa 4.1 että kuvassa 4.1 on esitetty esimerkinomaisesti kuluttajabarometrin vastaajien jakautuminen ikäryhmittäin alkuvuodelta 2019.



Kuva 4.1: Vastaajien jakautuminen ikäryhmittäin tammi–huhtikuu 2019

Ikä	15–24	25–34	35–44	45–54	55–64	65–74	75–84	yhteensä
vastattu	9,5	12,7	14,6	15,4	**17,8**	**19,9**	**10,1**	100 %
ei vast.	**16,6**	**18,6**	16,4	15,4	14,3	11,9	6,8	100 %
otos	13,0	15,7	15,5	15,4	16,0	15,9	8,0	100 %

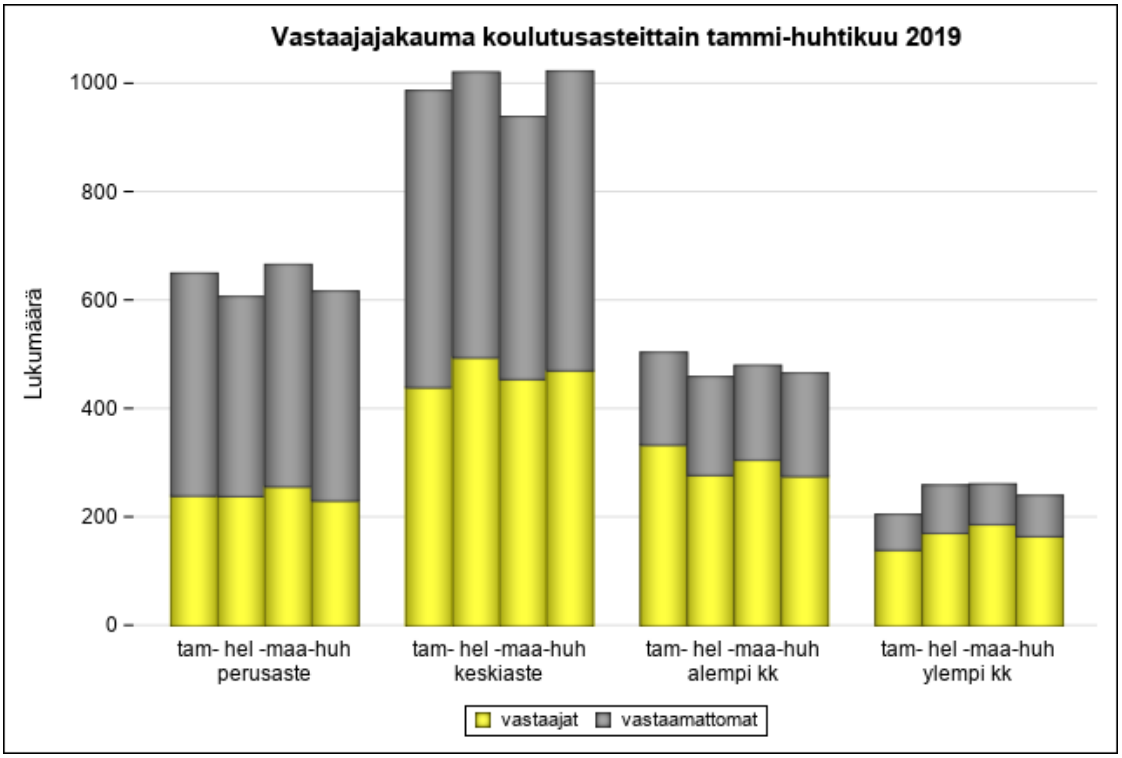
Taulukko 4.1: Vastaajien jakautuminen ikäryhmittäin tammi–huhtikuu 2019

Taulukon 4.1 alariviltä näemme kuinka ikäluokat jakautuvat täydessä otoksessa, eli ne osuudet, jotka vastanneilla tulisi olla. Sekä taulukosta että kuvaajasta havaitaan, että vastausaste kasvaa iän kasvaessa. Vastaavasti taulukosta 4.2 ja kuvasta 4.2 nähdään samat tiedot koulutuksen osalta. Tällä kertaa havaitaan, että vastausaste kasvaa koulutustason kasvaessa. Lisäksi kuvaajista havaitaan, että kuukausien välillä ei ole merkittävää eroa kummassakaan luokituksessa. Vastaavat kuvaajat neljän seuraavan kuukauden ajalta yhdistelmätiedonkeruun osalta löytyvät kappaleesta 4.2.1.

Ikä	perusaste	keskiaste	alempi kk.	ylempi kk.	yhteensä
vastattu	20,6	39,7	**25,5**	**14,2**	100 % / 4685
ei vast.	**33,4**	**44,8**	15,2	6,5	100 % / 4715
otos	27,1	42,3	20,4	10,3	100 % / 9400

Taulukko 4.2: Vastaajien jakautuminen koulutusasteittain tammi–huhtikuu 2019

Sukupuolen ja alueiden jakaumat ovat tässä suhteessa niin tasaisia, että ne voidaan sivuuttaa. Esimerkiksi miesten vastausaste vuoden 2019 neljän ensimmäisen kuukauden aikana vaihtelee välillä $(0,46–0,51)$ ja naisten vastaavasti välillä



Kuva 4.2: Vastaajien jakautuminen koulutusasteittain tammi–huhtikuu 2019

(0, 49–0, 52). Alueiden vastausasteet vaihtelevat samalla ajanjaksolla välillä (0, 47–0, 52). Näiden kahden muuttujan vaihtelu on suunnilleen samansuuruista koko tutkimusaikavälillä.

4.1.2 Logistinen regressiomalli

Jotta R-indikaattorien arvoja voi vertailla keskenään, niiden laskentaan käytettyjen muuttujien ja luokitusten tulee olla samat. Täytyy siis löytää kaikille aineistoille yhteiset merkitsevät selittäjät vastausalttiuden estimointiin. Aineistosta löytyvät taustamuuttujat, joille siis löytyy arvo sekä vastanneille että vastaamattomille, ovat: ikä, suuralue, koulutus ja sukupuoli. Näistä muuttujista ainoastaan dikotomiseksi luokiteltu sukupuoli käy sellaisenaan malliin selittäjäksi. Ikä suhdeasteikollisena muuttujana toimisi sellaisenaan logistisen regressiomallin puolesta, mutta R-indikaattoria laskettaessa se tulee uudelleen luokitella, jotta kaikkiin luokkiin saadaan mielekäs määrä havaintoja. Ikäluokkia on joko 6 tai 7 riippuen siitä tutkitaanko kuluttajabarometriaineistoa vai kuluttajien luottamus -aineistoa. Kuten aiemmin on mainittu, sekä järjestysasteikollinen koulutus että laatueroasteikollinen suuralue, täytyy uudelleen luokitella dikotomisten indikaattorimuuttujien joukoiksi. Näiden luokittelujen jälkeen mahdollisia selittäviä muuttujia jää 11 kappaletta:

- Ikäluokka: 6 / 7 luokkaa (ikal)
- sukupuoli: 2 luokkaa (sp)
- koulutusklusteri: kaksiluokkaisia indikaattorimuuttujia

  - perusaste (tutpe)
  - keskiaste (tutke)
  - alempi korkea-aste (tutak)
  - ylempi korkea-aste (tutyk)

- alueklusteri: kaksiluokkaisia indikaattorimuuttujia
  - pääkaupunkiseutu (AlueH)
  - Etelä-Suomi (AlueE)
  - Länsi-Suomi (AlueL)
  - Itä-Suomi (AlueI)
  - Pohjois-Suomi (AlueP)

Edellä mainittujen selittäjien lisäksi mallissa on mukana vakiotermi. Selitettävä muuttuja on tässä tapauksessa *respons*, eli se onko henkilö vastannut kyselyyn vai ei. Tarkemmin ottaen mallinnetaan tilannetta, että henkilö ei ole vastannut kyselyyn, eli että *respons* saa arvon 0.

Koska kuukausiaineistoja on vähimmilläänkin neljä, mutta enimmillään peräti 92, niin kaikkien mahdollisten mallien testaus ei ole mielekästä. Käytetään siis sopivan selittäjäkombinaation löytämiseksi viimeistä kahtatoista kuukautta. Merkitsevän selittäjäkombinaation löytämiseksi valitaan askeltava (stepwise) muuttujien valintamenettely. Askeltavan menetelmän tiedetyistä ongelmista, kuten siitä että menetelmä ei mahdollisesti löydä optimaalista mallia, ei pitäisi tässä suhteellisen selkeässä tilanteessa olla haittaa. Haitoilta suojaa myös se, että tässä haetaan yhteistä merkitsevää selittäjäjoukkoa useille aineistoille, jolloin yksittäisten kuukausiaineistojen poikkeavat tilanteet voidaan jättää huomiotta.

Kahdentoista kuukauden aineistojen mallinnuksen perusteella tulokseksi saadaan, että ikäluokka on ainoa kaikkina kuukausina merkitsevä selittäjä. Taulukosta 4.3 näemme myös, että koulutusluokista perusaste on merkitsevä yhdeksänä kuukautena, keskiaste kahdeksana, alempi korkea-aste seitsemänä ja ylempi korkea-aste kuutena kuukautena. Sukupuoli on merkitsevä selittäjä viitenä kuukautena. Aluemuuttujista Länsi-Suomi on merkitsevä neljänä kuukautena, pääkaupunkiseutu kahtena kuukautena ja muut eivät ollenkaan. Valintakriteerinä käytetään Waldin testiä ja merkitsevyystasoksi valitaan 5 prosenttia.

Tällä perusteella selittäjien joukosta pudotetaan pois sukupuoli ja aluemuuttujat. Koulutusmuuttujista yksi jätetään pois automaattisesti, koska kyse on indikaattorimuuttujaryhmästä, sillä jokainen ryhmän yksittäinen muuttuja on aina ryhmän muiden muuttujien lineaarikombinaatio. Malliin valitaan siis muuttujat ikäluokka ja koulutus. Vakio ei ole merkitsevä kahtena kuukautena, mutta siitä huolimatta se jätetään malliin. Lisäksi tulee vielä selvittää, onko muuttujilla yhdysvaikutusta. Kun koulutuksen ja ikäluokan yhdysvaikutus lisätään malliin, niin se on testatun vuoden aikana yhtenä kuukautena merkitsevä. Yhdysvaikutustermi voidaan siis jättää pois mallista.

Kun tällä valikoituneella mallilla tehdään samanlainen kahdentoista kuukauden testiajosarja, niin havaitaan, että ikäluokka ja koulutus ovat molemmat kaikkina

kk	vakio	ikal	perus	keski	alempi	ylempi	SP	Länsi	PKS
7	<,0001	<,0001	<,0001	<,0001					
8	<,0001	<,0001			<,0001	<,0001			0,0369
9	<,0001	<,0001	<,0001	<,0001			0,0041		
10	0,0595	<,0001	0,0262		<,0001	0,0007			
11	<,0001	<,0001	<,0001	<,0001			0,002	0,0003	
12	<,0001	<,0001		0,0053	<,0001	<,0001			
1	<,0001	<,0001	0,0004		<,0001	<,0001	0,0359	0,0004	
2	<,0001	<,0001	<,0001	<,0001				0,0371	
3	<,0001	<,0001	<,0001	<,0001	0,0168		0,0145	0,0168	
4	<,0001	<,0001		0,0006	<,0001	<,0001			0,0012
5	0,3404	<,0001	0,0493		0,0111	<,0001	0,0018		
6	<,0001	<,0001	<,0001	<,0001					
lkm	10	12	9	8	7	6	5	4	2

Taulukko 4.3: Logistisen regressiomallin merkitsevien selittäjien p-arvot ajanjaksolla 7/2018 – 6/2019

kuukausina merkitseviä selittäjiä. Toisaalta, kun selittäviä muuttujia käsitellään luokittelevina muuttujina, niin havaitaan, että osa ikäluokista ei ole jatkuvasti merkitseviä. Ikäluokat $15/18 - 24$, $45 - 54$ ja $55 - 64$ osoittautuivat useimmiten ei-merkitseviksi selittäjäluokiksi. Tästä huolimatta ikäluokka muuttujaa käytetään sellaisenaan selittäjänä, koska se on kokonaisuudessaan kuitenkin merkitsevä. Koulutusmuuttujan kaikki luokat säilyvät merkitsevinä selittäjinä kaikkina kuukausina.

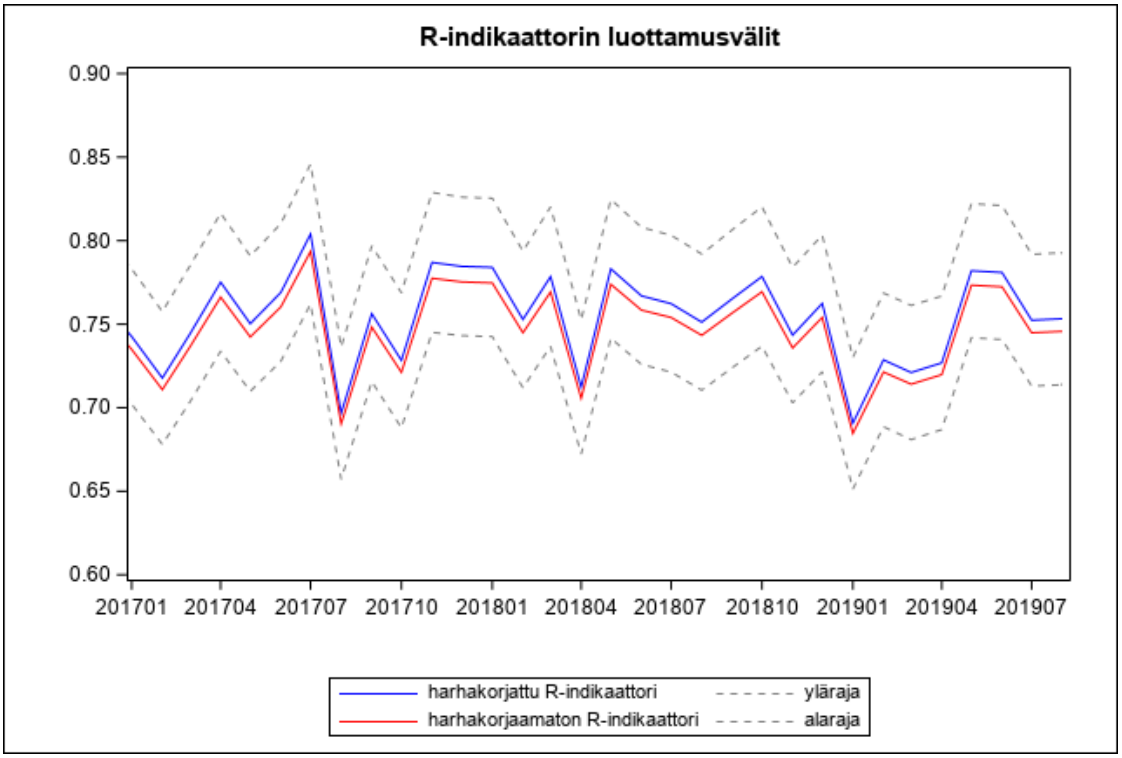
Lopuksi on syytä todeta, että RISQ-projektin SAS-koodin logistisessa regressiomallissa kaikki muuttujat määritellään luokitteleviksi muuttujiksi, mikä ei sinällään olisi pakollista R-indikaattorin laskennan kannalta. Pääsyy tähän valintaan lienee se, että osittaisten R-indikaattorien laskennassa tällainen luokittelu on tarpeen. Tällä valinnalla toki samalla varmistuu, etteivät käyttäjät riko muuttujien mitta-asteikkovaatimuksia. Valinnalla on vaikutusta kaikkiin niihin tuloksiin, joiden laskennassa käytetään logistisella regressiomallilla estimoitua vastausalttiutta ρ. Luokittelun vaikutuksesta R-indikaattoriin saadaan jonkinlainen käsitys taulukon 4.4 perusteella. Taulukosta nähdään neljän viimeisen kuukauden R-indikaattorit luokitellun ja luokittelemattoman aineiston pohjalta laskettuna. Taulukosta havaitaan, että tehty luokittelu pienentää R-indikaattorin arvoa. Luokituksen aiheuttama ero vastausalttiuksissa ei tule esiin taulukossa käytetyllä tarkkuudella.

R-indikaattori	luokiteltu	luokittelematon	vastausalttius
toukokuu	0,782	0,814	0,470
kesäkuu	0,781	0,793	0,463
heinäkuu	0,752	0,762	0,477
elokuu	0,753	0,757	0,491

Taulukko 4.4: Luokittelun vaikutus R-indikaattorin arvoon

4.1.3 R-indikaattorin laskenta

Tässä tutkielmassa R-indikaattorin laskentaan käytetään kahta eri SAS-koodia. RISQ-projektin lopullista SAS-koodia käytetään tässä tutkielmassa osittaisten R-indikaattorien laskennassa. Lisäksi tällä SAS-koodilla määritellään R-indikaattorin luottamusvälit ja lasketaan R-indikaattorin harhakorjattu versio (kuva 4.3). Harhakorjattu ja harhakorjaamaton R-indikaattori eivät poikkea merkittävästi toisistaan. Harhakorjattu R-indikaattori on jatkuvasti aavistuksen korjaamatonta pienempi ja mitään eroa kuvaajien profiileissa ei ajan edetessä näy. Edellä mainitun vuoksi harhakorjausta ei erityisesti huomioida jatkotarkasteluissa. RISQ-projektin alkuperäisessä SAS-koodissa hajontaa estimoitiin bootstrap-menetelmällä, kun taas lopullisessa koodissa hajonta estimoidaan analyyttisen kaavan pohjalta. [de Heij (2010)]



Kuva 4.3: R-indikaattorin luottamusvälit sekä harhakorjaamaton R-indikaattori

Tätä tutkielmaa varten laadittiin kolme eri SAS-koodia R-indikaattorien laskentaan. Ensimmäisessä ei estimoitu hajontaa lainkaan. Tämä koodi tuotti laskentatarkkuuden puitteissa samat tulokset kuin RISQ-projektin varsinainen koodi. Seuraavassa versioissa lisättiin hajonnan estimointi bootstrap-menetelmällä. Tällä koodilla hajonta ei konvergoitunut järjellisillä toistomäärillä ja itse R-indikaattorin estimaattikin oli alaspäin harhainen. Varsinaista syytä tähän ei löytynyt, mutta luultavimmin ongelmat johtuivat R-indikaattorin monivaiheisesta laskennasta. Koska logistisella regressiomallilla estimoidut vastausalttiudet (ρ) vaihtelivat suuresti, niin päätettiin siirtyä vakaampaan Jackknife-menetelmään. Jackknife-menetelmä tuottikin odotetusti vakaampia vastausalttiuksia. Tällä menettelyllä sekä itse R-indikaattori että saadut hajontaestimaatit olivat mielekkäitä. Koodin ajoaika oli kuitenkin

selkeästi pidempi, kuin alkuperäisellä RISQ-projektin analyyttisellä koodilla, joten loppujen lopuksi päädyttiin käyttämään sitä.

Näin ollen tutkielmaa varten laadituista SAS-koodeista käytetään lopulta vain ensimmäistä (Liite 6.1.1), sillä sillä lasketaan samalla kaikki vaihtoehtoiset indikaattorit. Tutkielmaa varten laaditussa SAS-koodissa ei korvata ikäluokkamuuttujaa indikaattorimuuttujilla, koska tämä ei ole tarpeen ja samalla saadaan vertailuun vielä yksi hieman eroava malli. Näiden mallien vastausalttiusestimaatit poikkeavat toisistaan hyvin vähän, mutta niiden avulla lasketut R-indikaattorit poikkeavat kuitenkin selkeästi toisistaan (ks. edellä esitetty taulukko 4.4).

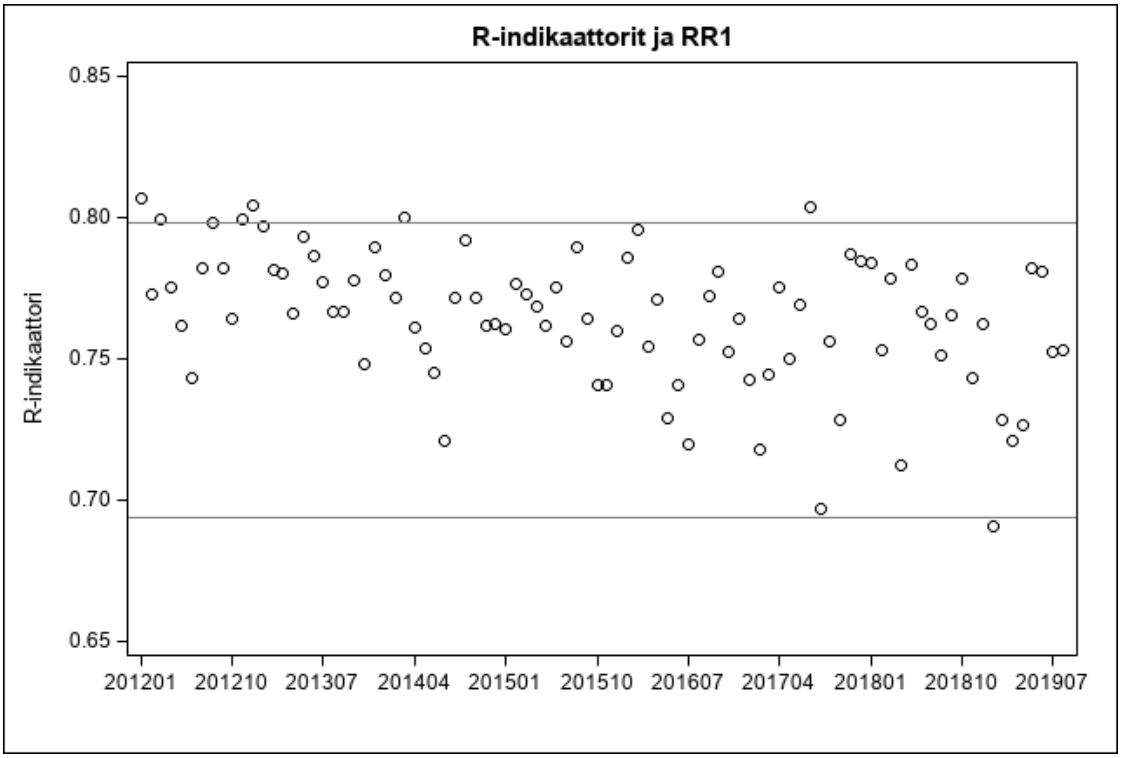
4.1.4 R-indikaattorin tulkinta

Artikkelissa *Representativeness indicators for measuring and enhancing the composition of survey response*, Schouten ja Bethlehem (2009) määrittävät R-indikaattorille kolme käyttötarkoitusta:

1. Samasta perusjoukosta tehtyjen tutkimusten vertailu
2. Saman tutkimuksen vertailu ajan kuluessa
3. Aineiston keruuvaiheessa tehty tarkkailu

Mikään näistä ei suoraan sisällä tapaa tulkita yksittäisen R-indikaattorin arvoa. Artikkelissa kuitenkin määritellään kolme mittaria: RR_1, RR_2 ja RR_3, joita voidaan pitää jonkinlaisina hyvyysmittareina. Mittarit on määritelty luvussa 3.2.8 *Tulkinta*.
Näistä RR_1 määrittää alarajan R-indikaattorin hyvyydelle. Hyvyysraja määrittyy seuraavasti: Vaaditaan, että (1-α):n määrittelemä osuus vastausalttiuksista on korkeintaan (1-γ):n etäisyydellä vastausasteesta. Artikkelissa käytetään 5 %:n merkitsevyystasoa, joka johtaa siihen, että vertailun voi käytännössä tehdä valittuun (1-γ):n arvoon. Eli kun kaavassa 3.33, $\zeta_{(1-\alpha)}$ paikalle sijoitetaan $1,960$, niin $RR_1 \approx 1 - \frac{2}{1,960}\gamma$, joten γ:n kertoimeksi jää noin $1,02$. Kun palautetaan mieleen, että R-indikaattorin arvot tutkimusaineistossa vaihtelevat välillä $0,691$–$0,807$, niin RR_1:sen arvojen tulisi olla suunnilleen vastaavalla tasolla. Jotta näin olisi, niin 5 %:n merkitsevyystasoa käytettäessä pitäisi valittu kynnysarvo γ asettaa parhaimmassakin tilanteessa noin 20 prosenttiin. Kuvassa 4.4 nähdään R-indikaattorin arvot ja RR_1:n määrittämät hyvyysrajat, kun kynnysarvona (γ) käytetään 20 ja 30 prosenttia. On siis selvää, että edes tavoiteltua minimitasoa ($R > 0,796$) ei tässä aineistossa saavuteta kuin satunnaisesti.

RR_2 on hyvyysmittari, joka on johdettu R-indikaattorin harhan $B_m(X)$:n kaavasta. Kaavasta 3.34 voidaan havaita, että RR_2:n arvoon vaikuttaa γ:n arvon lisäksi vastausalttius, joka tutkimusaineistossa vaihtelee välillä $0,45$–$0,65$. Kun vastausaste oli yli 60 prosenttia, niin aineistosta lasketut R-indikaattorit osuivat aina välille, jossa γ:n arvot ovat 10 ja 20 prosenttia. Kun vastausasteet laskivat alle 50 prosentin, niin R-indikaattorit eivät enää pysyneet edes näiden hyvyysrajojen puitteissa. Ajan kuluessa ja vastausasteen laskiessa harhattomuuden osalta tilanne muodostui yhtä huonoksi kuin edustavuudenkin.



Kuva 4.4: R-indikaattorin arvot ja RR_1:n hyvyysrajat, kuin γ on 20 ja 30 prosenttia.

RR_3 on johdettu havaittujen ja havaitsemattomien vastausten kontrastista. Kaavasta 3.35 nähdään, että vaikuttavat tekijät ovat tässäkin hyvyysmittarissa vastausalttius ρ ja kynnysarvo γ. Kolmesta hyvyysmittaristä tällä on jatkuvasti korkeimmat arvot, joten tutkimusaineistosta lasketut R-indikaattorit eivät näitä hyvyysrajoja saavuta. Jos γ:n arvoksi asetetaan 20 prosenttia, niin RR_3:n arvo on tällöin jo yli 90 prosenttia ja γ:n arvoa pienennettäessä se siitä vain kasvaa.

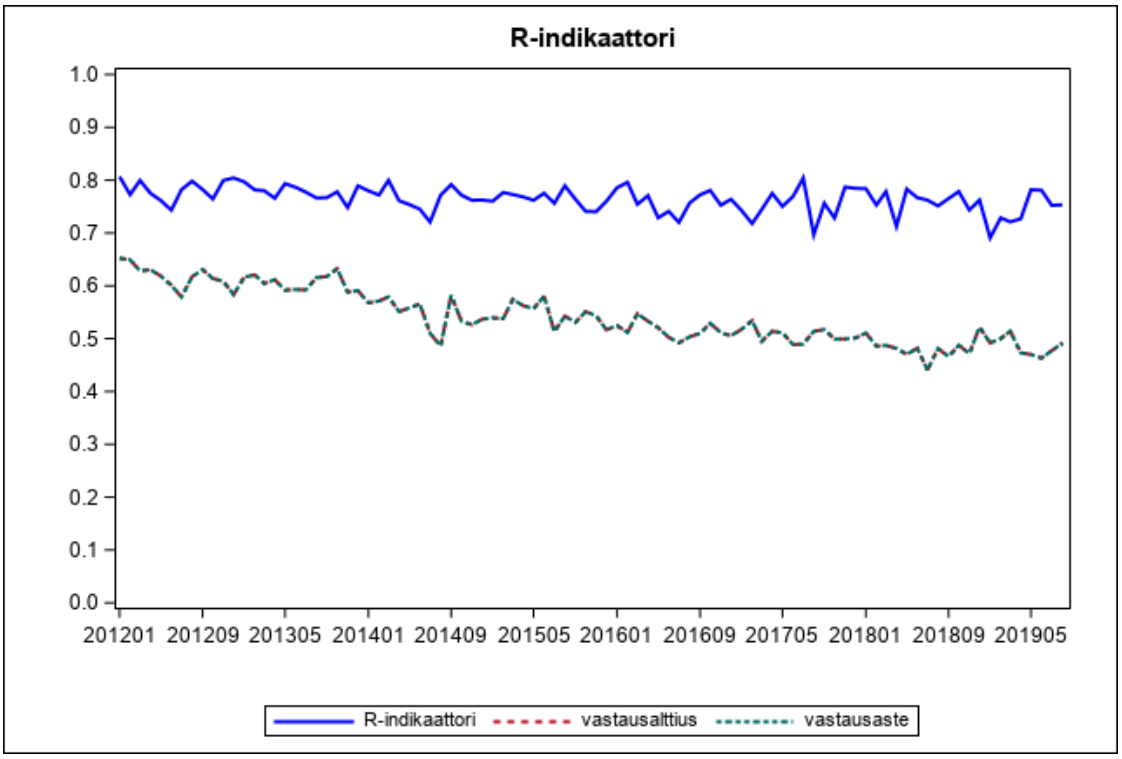
Taulukossa 4.5 on esimerkinomaisesti elokuun 2019 mukaiset vertailuarvot käytetyille hyvyysmittareille. RR_1:n osalta on taulukoitu sekä 5:n että 10:n prosentin merkitsevyystasot. Elokuun R-indikaattorin arvo on $0,753$ ja vastausaste on $0,491$. Havaitaan, että 5:n prosentin merkitsevyystasolla pitäisi käyttää kynnysarvona 30:tä prosenttia, jotta edustavuus saavuttaisi tämän rajan. Harhan osalta ollaan myös samalla 30:n prosentin tasolla. Kontrastin suhteen ei päästä edes 30:n prosentin kohdalla lähellekään vertailuarvoa.

kynnysarvot (γ)	0,05	0,10	0,20	0,30
RR_1 (edustavuus 5 %)	0,949	0,898	0,796	0,694
RR_1 (edustavuus 10 %)	0,939	0,878	0,757	0,635
RR_2 (harha)	0,951	0,902	0,804	0,705
RR_3 (kontrasti)	0,975	0,950	0,900	0,850

Taulukko 4.5: Hyvyysmittarien vertailuarvot elokuu 2019

4.1.5 R-indikaattorin kehitys ajassa

Kuvaajasta 4.5 nähdään, että ajan edetessä vastausasteet selkeästi laskevat. R-indikaattorin vastaava kehityssuunta ei ole kuvaajasta 4.5 aivan yhtä selkeästi havaittavissa. Kun sovitetaan aineistoon regressiomallit, joissa ajalla selitetään vastausasteen tai vaihtoehtoisesti R-indikaattorin muutosta, niin laskeva tendenssi nousee esiin (kuva 4.6).



Kuva 4.5: R-indikaattorin ja vastausasteen kehitys ajan kuluessa

Kun sovitetaan aineistoon lineaarinen regressiomalli, jossa vastausastetta selitetään ajalla (yhtälö 4.1), niin havaitaan, että vastausaste laskee noin 0,02 yksikköä vuodessa. Aika on mallissa merkitsevä selittäjä ja mallin selitysaste on korkea ($R^2 = 0,838$).

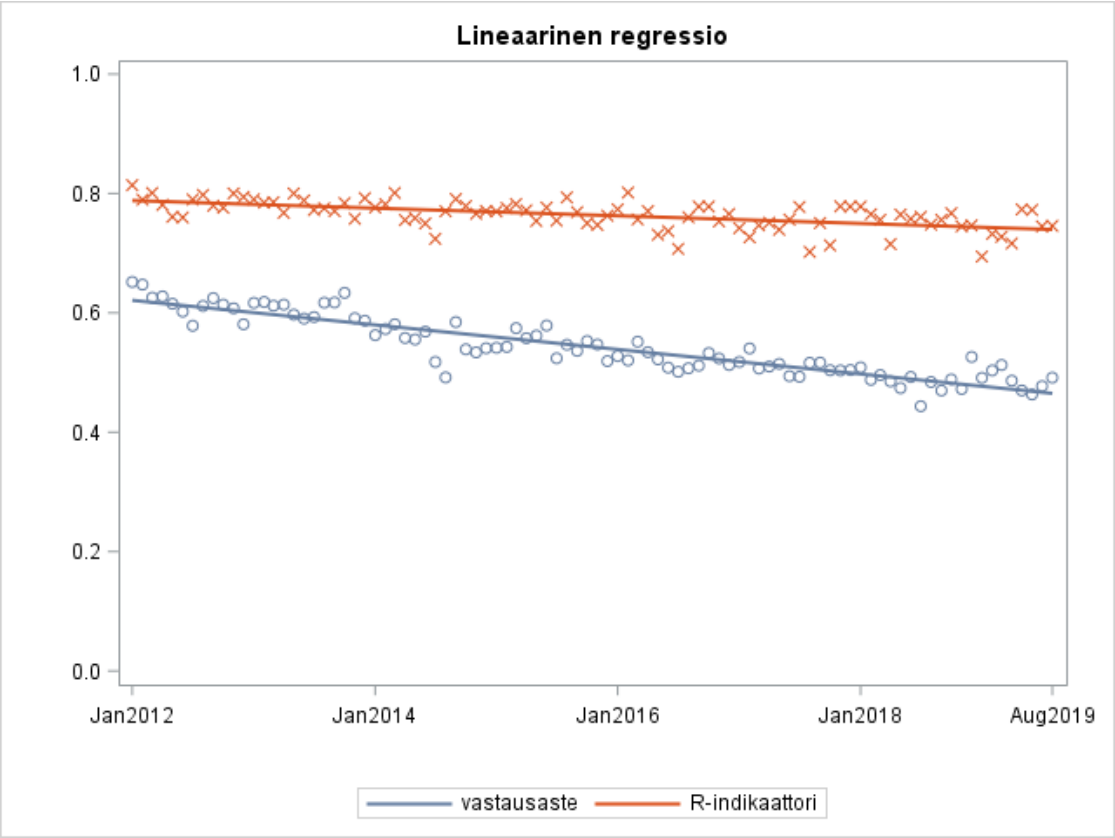
$$vastausaste = -0,0206 * aika + 0,621 \tag{4.1}$$

Kun vastaavasti R-indikaattorin muutosta selitetään ajalla (yhtälö 4.2), niin havaitaan, että myös R-indikaattorin arvot laskevat ajan edetessä. Lasku ei ole yhtä nopea kuin vastausasteen kohdalla, sillä R-indikaattori pienenee vain noin 0,006 yksikköä vuodessa. Mallin selitysaste on tässä selkeästi pienempi ($R^2 = 0,351$) kuin edellisessä.

$$R(\rho) = -0,0064 * aika + 0,788 \tag{4.2}$$

Käytetyssä mallissa (yhtälöt: 4.1 ja 4.2) aika on skaalattu vuosiksi, mutta laskentatarkkuutena käytetään kuitenkin kuukautta. Tässä mallissa ajan 0-kohdaksi on valittu tammikuu 2012. Eli siis kun aika lisääntyy vuodella, niin vastausaste laskee

keskimäärin 0,02 yksikköä. Ja vastaavasti R-indikaattori laskee keskimäärin 0,006 yksikköä.



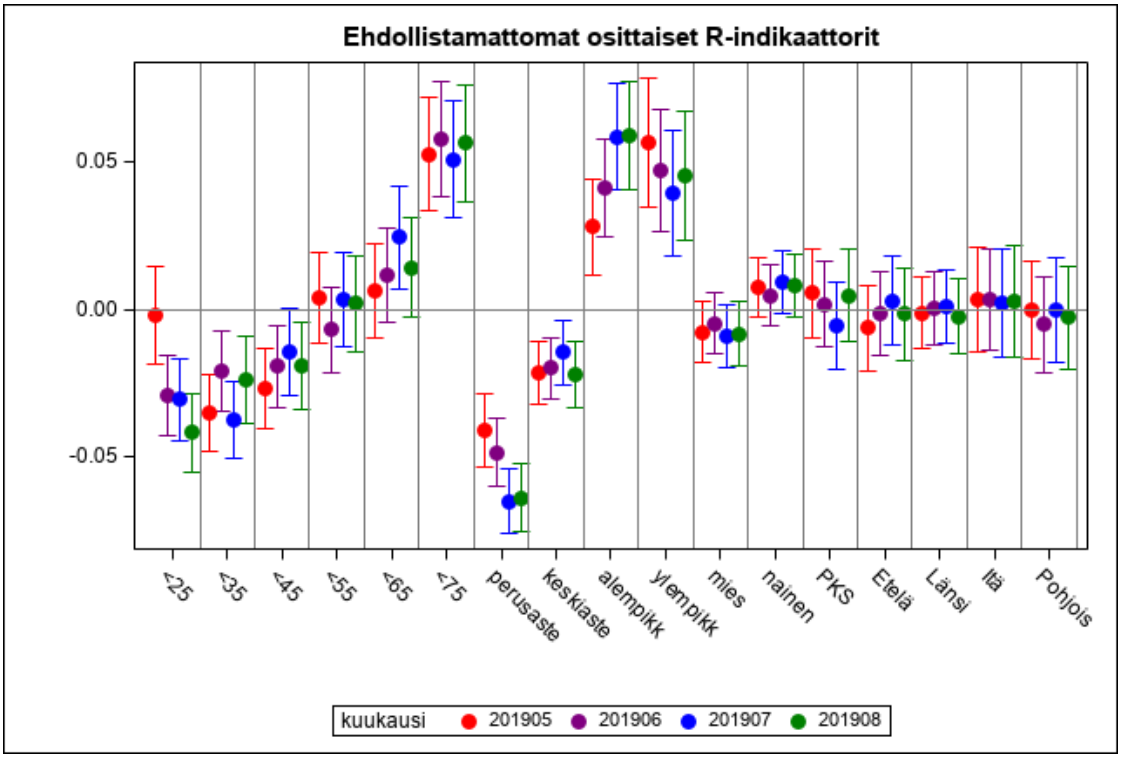
Kuva 4.6: Regressiosuorat, kun ajalla selitetään R-indikaattoria ja vastausastetta

Tehdyissä aikasarjatarkasteluissa havaittiin lineaarinen trendi, mutta ei havaittu kausivaihtelua. Tämän vuoksi päätettiin käyttää lineaarista regressiomallia ajallisessa tarkastelussa, varsinaisten aikasarjamallien sijaan.

4.1.6 Osittaiset R-indikaattorit

Osittaisten R-indikaattoreiden tarkoituksena on kertoa, mitkä luokat ovat yli- tai aliedustettuina otoksessa. Ehdollistamattomia osittaisia luokiteltuja R-indikaattoreita käytettäessähän voidaan tutkia sekä mallintamisessa käytettyjä (ikäluokka ja koulutusaste), että pois jätettyjä (sukupuoli ja alue) taustamuuttujia. Näissä tarkasteluissa aineiston ikäluokat ovat yhdistelmätiedonkeruun mukaiset, eli aiemmista aineistoista on poistettu ylin ikäluokka ja alimman ikäluokan alaraja on kahdeksantoista vuotta.

Kuvassa 4.7 on yhdistelmätiedonkeruun aikaiset kuukaudet (touko–kesäkuu). Kuvasta on selkeästi nähtävissä miksi ikäluokka ja koulutusaste valikoituivat mallin selittäviksi muuttujiksi ja miksi sukupuoli ja alue voitiin jättää mallista pois. Kuvasta nähdään myös, että nuoremmat ikäluokat ovat aliedustettuina ja vanhemmat yliedustettuina vastanneiden joukossa. Samaten on nähtävissä, että koulutusasteen



Kuva 4.7: Osittaiset *R*-indikaattorit yhdistelmätiedonkeruussa

kasvu nostattaa aluksi vastausalttiutta, mutta ylemmän ja alemman korkean asteen koulutuksen välillä ei eroa enää ole.

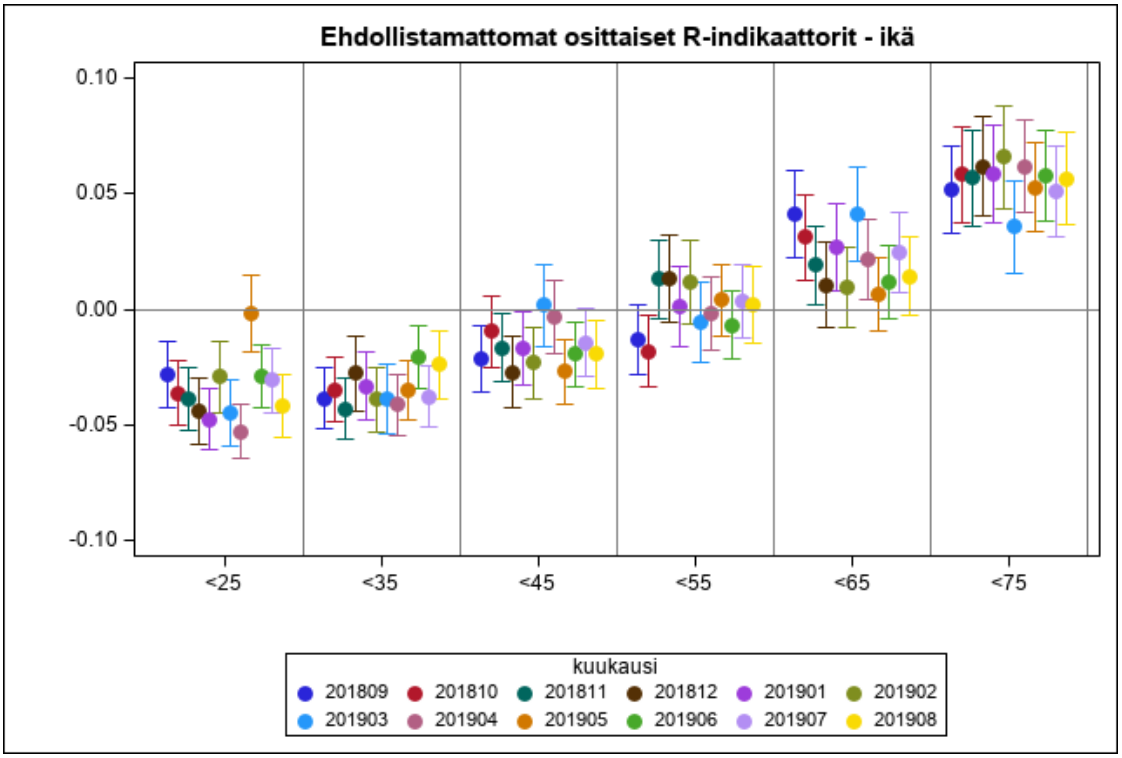
Tarkastellaan seuraavaksi hieman tarkemmin yksittäisten muuttujien tilannetta hieman pidemmällä aikavälillä, jotta saadaan käsitys siitä, onko siirtymisellä yhdistelmätiedonkeruuseen ollut vaikutusta vastausalttiuteen.

Vuoden mittaisella tarkastelujaksolla näyttäisi siltä, että nuorimman ikäluokan kohdalla siirtyminen yhdistelmätiedonkeruuseen on hieman parantanut tilannetta (kuva 4.8). Muissa ikäluokissa ei voi havaita merkittäviä muutoksia. Muutos nuorimmassa ikäluokassa saattaa toki johtua 15–17 -vuotiaiden poistumisesta tai kausivaihtelusta. Näihin kysymyksiin palataan tarkemmin luvussa 4.2 *Yhdistelmätiedonkeruu*.

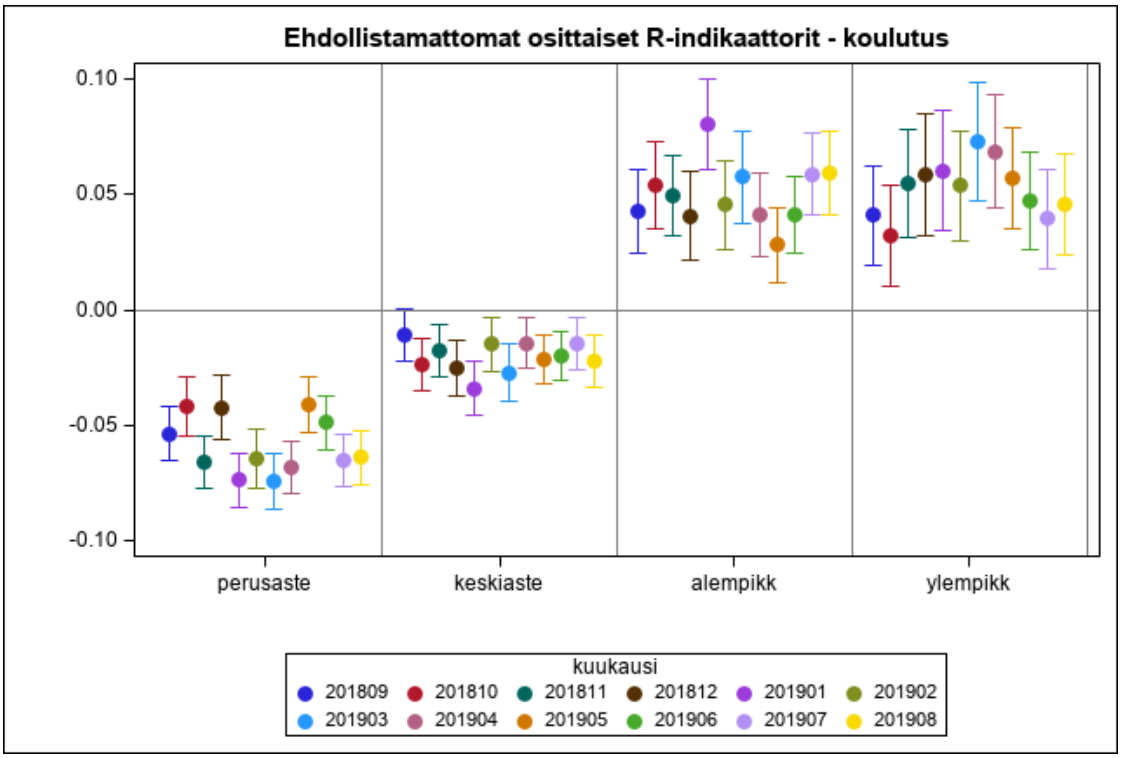
Kuvan 4.9 perusteella koulutusasteen osalta ei ole nähtävissä systemaattisia muutoksia suuntaan tai toiseen, eli voitaneen sanoa, että siirtymästä ei ole ainakaan ollut haittaa. Kuvassa 4.7 näkyneet laskevat ja nousevat trendit näyttäisivät vuoden tarkastelujaksolla olevan osa luontaista vaihtelua. Näitäkin muutoksia tarkastellaan hieman tarkemmin luvussa 4.2 *Yhdistelmätiedonkeruu*.

Muuttujissa sukupuoli ja alue ei vuoden tarkastelujaksolla nouse esiin mitään aiemmasta kuvasta poikkeavaa. Naisten vastausalttius on jatkuvasti hieman miesten vastausalttiutta korkeampaa ja eri alueilla vastausalttiuden vaihtelu on varsin tasaista.

Herää kuitenkin kysymys, luoko osittaisten *R*-indikaattorien laskeminen lisäinformaatiota suhteessa luokittaisten vastausalttiuksien vaihtelun tarkasteluun. Jos esimerkiksi tarkastellaan elokuuta 2019, niin vanhin ikäluokka on n. 13 prosenttiyksikköä yliedustettuna ja vastaavasti alin ikäluokka on n. 13



Kuva 4.8: Osittainen R-indikaattori – ikäluokka



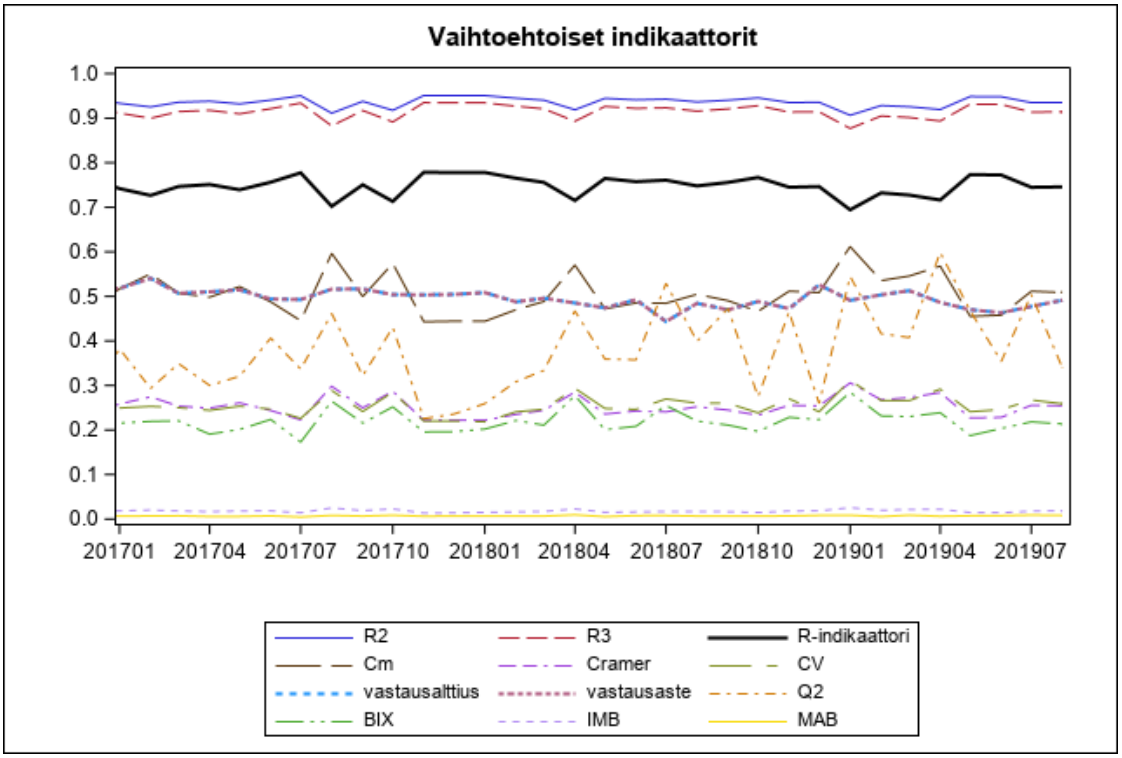
Kuva 4.9: Osittainen R-indikaattori – koulutusaste

prosenttiyksikköä aliedustettuna. Väliin jäävät ikäluokat poikkeavat merkittävästi vähemmän toisistaan. Tulokset siis näyttävät hyvinkin samansuuntaisilta, kuin

osittaisista R-indikaattoreista on pääteltävissä (kuva 4.7). Myös alimassa ikäluokassa havaittava poikkeava käytös toukokuulta näkyy vastausasteissa hyvin lähellä nollaa olevana poikkeamana, joka vastaa kuvaajan käytöstä .

4.1.7 Vaihtoehtoiset indikaattorit

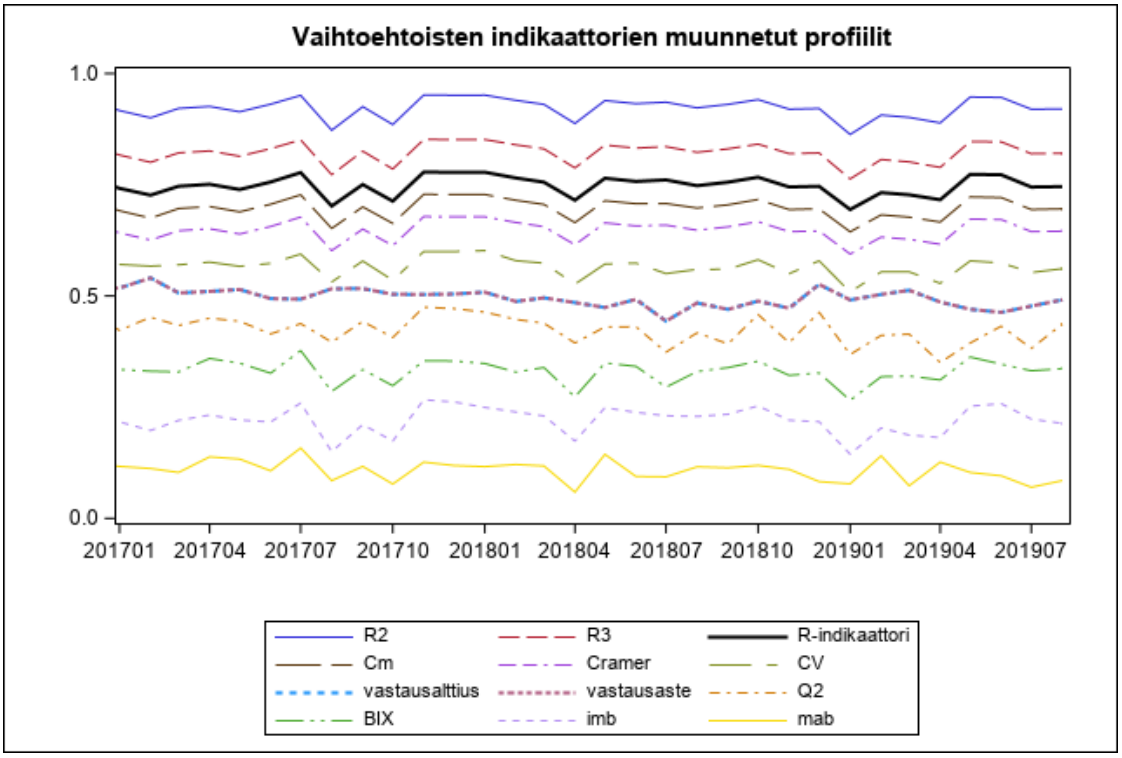
Vaihtoehtoisten indikaattorien vertailu suoraan on hieman vaikeaa, kun niiden skaalat ovat kovin erilaisia, kuten kuvasta 4.10 näkyy. Koska kaikkien vertailtavien indikaattorien arvo on kuitenkin vähintään nolla ja maksimissaan yksi, niin niiden yhdenmukaistaminen on helpohkoa. Samankaltaisuudet saadaan kuvaajatasolla näkyviin, kun tehdään tarvittavat lineaarimuunnokset. Pääasiassa siis skaalataan muuttujia, jotta vaihtelu saadaan näkyviin. Lisäksi joissain tapauksissa täytyy vielä vähentää indikaattorin arvo yhdestä, jotta kuvaajien nousut ja laskut ovat samansuuntaisia. Huomautettakoon vielä R_3-indikaattorin osalta, että riippumatta laskennassa käytetystä pseudo-$\Re^2$-tunnusluvusta (kaava: 3.50) päätelmät ovat samoja, ainoastaan käytetty lineaarimuunnos hieman muuttuu.



Kuva 4.10: Vaihtoehtoiset indikaattorit ja vastausaste

Kuvasta 4.11 nähdäänkin, että viisi ensimmäistä indikaattoria (eli kolme R-indikaattoria, C_m ja Cramérin V) ovat käytännössä lähes toistensa lineaarimuunnoksia. CV poikkeaa hiukan, mutta on hyvin samankaltainen edellisten kanssa. CV:n keskeisin poikkeama heinäkuussa 2018, selittyy puhtaan matemaattisesti vastausasteen merkittävällä pudotuksella. CV:n laskentakaavassahan on nimittäjässä vastausalttiuden keskiarvo, joka selittää CV:n erot muuhun R-indikaattoriryhmään. Sama ilmiö on nähtävissä helmi–maaliskuussa 2017.

Loput neljä indikaattoria poikkeavat enemmän sekä R-indikaattorista että toisistaan. Suurimmat nousut ja laskut ovat kaikilla indikaattoreilla selkeästi huomattavissa. BIX-indikaattori ja IMB-estimaattori kulkevat pääosin pääryhmän viitoittamaa tietä, vaikka molemmilla on pienet poikkeamansa. MAB-indikaattorilla on aivan oma lasku ja nousu maalis–huhtikuussa 2019. Tämä on hieman outoa, kun luulisi järjestystunnuslukuihin perustuvan indikaattorin olevan kaikkein vakain käyttäytymisessään. Eniten poikkeava vaihtoehtoinen indikaattori näyttäisi olevan Q_2, jolla on useampia omia poikkeamia. Osalla matkaa Q_2 näyttäisi olevan vakaampi ja toisaalla taas vaihtelevampi kuin muut indikaattorit.



Kuva 4.11: Muokatut indikaattoriprofiilit. R-indikaattori ja vastausaste/vastausalttius ovat muokkaamattomia, muut indikaattorit on skaalattu tarkoitushakuisesti.

Kuvassa 4.11 ainoastaan vastausaste/vastausalttius ja alkuperäinen R-indikaattori ovat muokkaamattomia, joten muiden indikaattorien osalta ei kannata kiinnittää huomiota kuin kuvaajan profiiliin. Nämä indikaattoreiden skaalaukset on tehty silmämääräisesti erilaisia kertoimia kokeillen ja lopuksi sovitettu, niin että ne eivät kuvassa risteä toistensa kanssa.

Tarkastellaan vielä R-indikaattorin ja muiden indikaattorien välistä korrelaatiota. Kun muuttujat asetetaan Pearsonin korrelaatiokertoimien itseisarvojen mukaiseen suuruusjärjestykseen, niin havaitaan, että kuvaajien perusteella tehdyt päätelmät osoittautuvat pääosin oikeiksi. Ainoa poikkeama on, että IMB-estimaattori näyttää korreloivan selkeästi vahvemmin R-indikaattorin kanssa kuin variaatiokerroin (CV). Kuvasta 4.11 näemme myös, että mikään indikaattoreista ei seuraa kaikkia vastausasteen muutoksia, eli niillä kaikilla on tässä mielessä jotain omaa kerrottavaa lähtötilanteeseen nähden. Indikaattorien korrelaatio vastausasteen kanssa puhuu

indikaattorit	R_2	V	R_3	Cm	IMB	CV	BIX	MAB	Q_2
R-indikaattori	0,998	-0,994	0,987	-0,973	-0,951	-0,839	-0,787	-0,652	-0,617
vastausaste	0,343	-0,246	0,210	-0,139	-0,360	-0,791	-0,764	-0,586	-0,833

Taulukko 4.6: Indikaattorien ja vastausasteen väliset korrelaatiot

selkeästi samaa kieltä (taulukko 4.6). Näemme, että Q_2:lla ja BIX:llä on kuitenkin selkeämpi yhteys vastausasteeseen, kuin muilla indikaattoreilla. Tämä selittyy sillä, että ne lasketaan otoskoon ja vastanneiden määrän perusteella. Toisaalta samoista lähtökohdista laskettu IMB-estimaattori ei kuitenkaan korreloi vahvasti vastausasteen kanssa. MAB:n korrelaatio löytyy edellisten väliltä. Toki myös CV:llä on korkea negatiivinen korrelaatio vastausasteen kanssa, joka selittyy kaavassa jakajana olevalla vastausalttiudella. Vastausasteen ja vastausalttiuden korrelaatiokerroin on laskutarkkuuden puitteissa 1, eli ne ovat käytännössä samat. Taulukon ulkopuolelta kerrottakoon vielä, että vastausasteen ja R-indikaattorin välinen korrelaatiokerroin on $0,340$. Samaisesta lähteestä, eli indikaattorien korrelaatiomatriisista, voidaan lisäksi todeta löytyvän korkea korrelaatio $(0,951)$ CV:n ja BIX:n väliltä.

4.1.8 Varsinaiset muuttujat

On ilmeistä, että tiedonkeruumenetelmän vaihtaminen vaikuttaa varsinaisten muuttujien jakaumiin, koska vastaajajakauma muuttuu. Kuluttajien luottamus -tutkimuksen keskiössä on neljän summamuuttujan muutokset ajassa (EU:n tasapainoluku ja kolme muuta indikaattoria). Summamuuttujille luontaiseen tapaan näihin vaikuttavat useiden muuttujien arvot, joten niiden vaihteluun vaikuttaa erittäin moni tekijä. Muistettakoon myös, että ennen tulosten laskentaa aineisto painotetaan vastaamaan perusjoukon oikeaa vastaajajakaumaa valittujen taustamuuttujien suhteen. Ja koska kyse on vastaajien mielipiteestä sen hetkisessä tilanteessa, niin luonnollisesti vastauksetkin muuttuvat sen hetkisen talous- ja elämäntilanteen mukaan.

Näin ollen on hyvin vaikea mitata tiedonkeruumenetelmän vaihtumisen aiheuttamaa muutosta suhteessa kaikkiin muihin tekijöihin. Jos aineistoa ei painotettaisi, niin jonkinlaisia vertailuja varsinaisten muuttujien jakaumista, voisi tehdä valittujen taustamuuttujien suhteen. Nytkin olisi mahdollista laskea painottamattomien muuttujien jakaumat ennen ja jälkeen tiedonkeruumenetelmän vaihtoa ja verrata niitä keskenään. Nämä tulokset kuitenkin poikkeavat lopullisista painotetuista tuloksista merkittävästi, joten mitään mielekästä tulkintaa niissä näkyville eroille ei pystytä antamaan ilman erittäin vahvaa substanssiosaamista. Tämän vuoksi näistä tarkasteluista on jouduttu luopumaan.

4.2 Yhdistelmätiedonkeruu

Yhdistelmätiedonkeruulla pyritään parantamaan sekä vastausalttiutta että otoksen edustavuutta. Odotusten mukaisesti nettikysely lisäsi nuorimpien ikäluokkien

vastausalttiutta. Vain vanhimmassa ikäluokassa nettivastaamisesta ei tullut suosituinta vastaamistapaa. Koska käytettiin yhdistelmätiedonkeruuta, niin vanhimmissakaan ikäluokissa vastausalttius ei pudonnut. Koulutusasteen osalta yhdistelmätiedonkeruu ei tasannut eroa vastausalttiudessa. Sukupuolen tai asuinalueen osalta ei tapahtunut mainittavia muutoksia.

4.2.1 Vastausaktiivisuus

Taulukosta 4.7 nähdään, että vastausasteella on pitkällä aikavälillä laskeva tendenssi. Vuonna 2012 vastausaste oli yli 60 prosenttia, kun se vuonna 2018 oli jo alle 50 prosenttia. Pyrkimys tämän tendenssin katkaisemiseen lienee keskeisin syy yhdistelmätiedonkeruuseen siirtymiselle. Muistutettakoon vielä mieleen, että tiedonkeruutavan muutoksen lisäksi samalla muutettiin mukaan otettuja ikäluokkia. Vastaajajoukosta poistui aktiivinen vastaajaluokka (75–84 v.) ikäjakauman yläpäästä ja vastaavasti passiivinen vastaajajoukko (15–17 v.) ikäjakauman alapäästä.

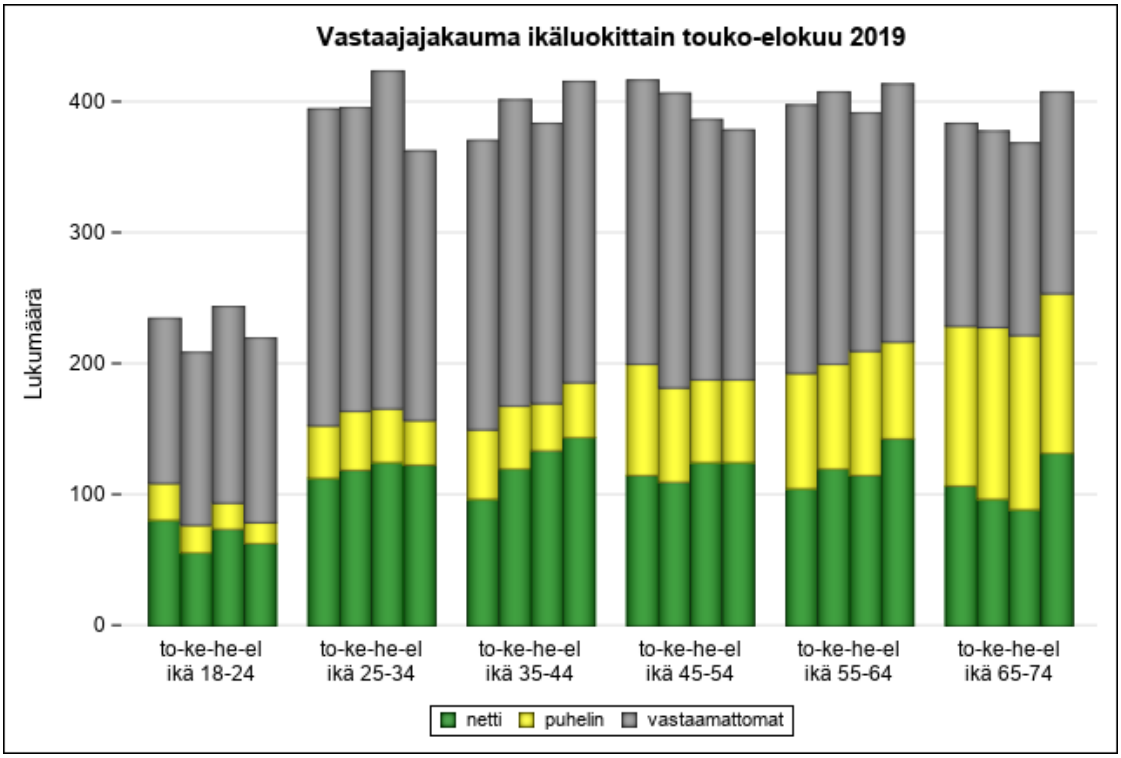
Vuosi	2012	2013	2014	2015	2016	2017	2018	2019p	2019y
vastausaste	0,616	0,607	0,550	0,548	0,521	0,510	0,486	0,498	0,475

Taulukko 4.7: Keskimääräiset vastausasteet vuosittain ja siirtymä puhelinhaastattelusta yhdistelmätiedonkeruuseen

Sekä taulukossa 4.8 että kuvassa 4.12 nähdään vastaamistapojen jakautuminen kuukausittain. Aineistosta havaitaan, että kaikissa ikäluokissa ja kaikkina kuukausina, kato on suurin luokka. Havaitaan myös, että vanhinta ikäluokkaa lukuun ottamatta nettikysely on yleisin vastaamistapa. Puhelinhaastattelu vastaamistapana on alimmissa ikäluokissa vähäinen, mutta kasvaa merkittäväksi haastattelutavaksi vastaajien iän noustessa. Mielenkiintoisena yksityiskohtana havaitaan, että viimeisenä kuukautena nettikysely oli suosituin vastaustapa myös vanhimmassa ikäluokassa.

	ikäluokka	18–24	25–34	35–44	45–54	55–64	65–74
toukokuu	kato	53,6	61,3	59,6	52,0	51,5	40,4
kesäkuu		63,2	58,6	58,2	55,3	51,0	39,7
heinäkuu		61,5	60,8	55,7	51,4	46,4	39,8
elokuu		64,1	56,7	55,3	50,4	47,6	37,7
toukokuu	puhelin	11,9	10,1	14,3	20,4	22,1	31,8
kesäkuu		10,0	11,4	11,9	17,7	19,6	34,7
heinäkuu		8,2	9,7	9,4	16,3	24,2	36,0
elokuu		7,3	9,4	10,1	16,6	17,9	29,9
toukokuu	netti	34,5	28,6	26,1	27,6	26,4	27,9
kesäkuu		26,8	30,1	29,9	27,0	29,4	25,7
heinäkuu		30,3	29,5	34,9	32,3	29,3	24,1
elokuu		28,6	33,9	34,6	33,0	34,5	32,4

Taulukko 4.8: Vastaustapajakauma prosentteina ikäluokittain



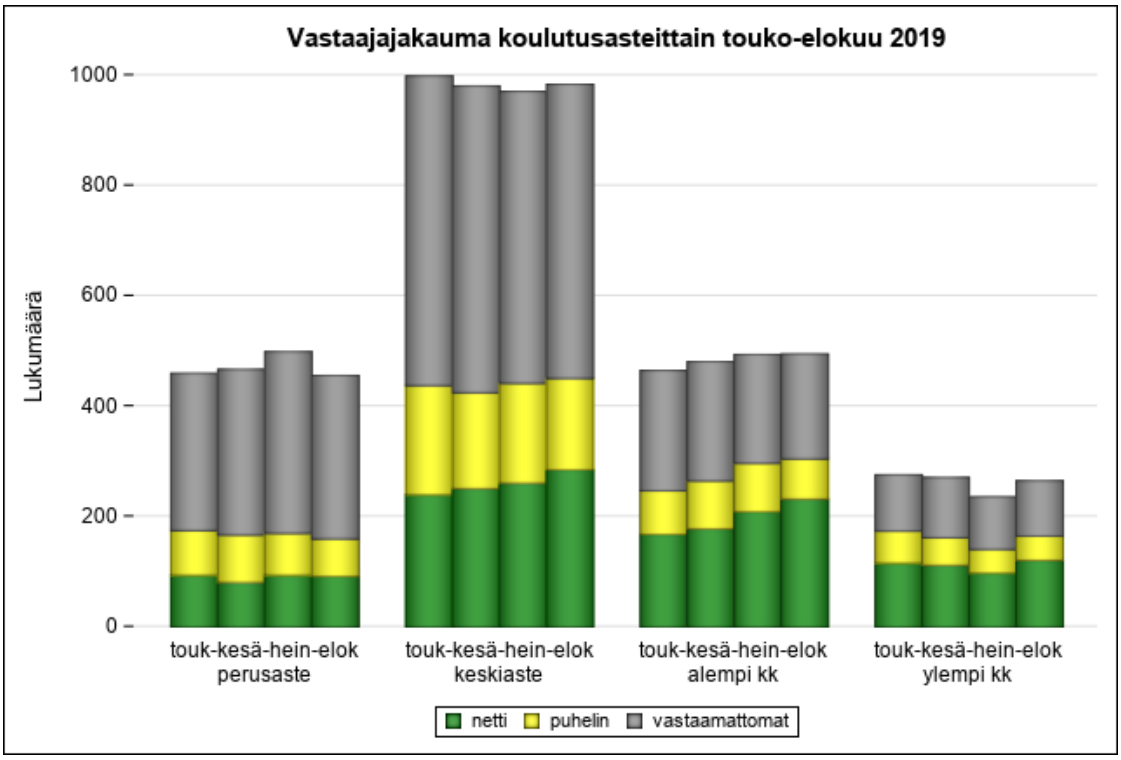
Kuva 4.12: Vastaustapa yhdistelmätiedonkeruun ajalta ikäluokittain

Taulukosta 4.9 ja kuvasta 4.13 voimme havaita, että koulutusasteittain luokiteltaessa nettivastaaminen on yhtä poikkeusta lukuun ottamatta yleisin vastaamistapa. Alimmassa koulutusluokassa netti- ja puhelinvastaaminen ovat sunnilleen samaa suuruusluokkaa. Mitä korkeampi koulutus vastaajalle on sitä todennäköisemmin hän vastaa kyselyyn netissä. Havaitaan myös, että koulutuksen lisääntyminen lisää vastaushalukkuutta. Korkeakoulutettujen vastausaste onkin yli 50 prosenttia. Pelkän perusasteen koulutuksen käyneillä vastausaste on alle 40 prosenttia. Keskiasteelle kouluttautuneet osuvat sitten edellä mainittujen väliin.

	koulutus	perusaste	keskiaste	alempi kk.	ylempi kk.
toukokuu	kato	62,0	56,2	46,9	37,0
kesäkuu		64,2	56,7	44,9	40,2
heinäkuu		65,9	54,5	39,9	40,3
elokuu		64,9	54,2	38,4	37,7
toukokuu	puhelin	17,6	19,8	17,0	21,0
kesäkuu		18,4	17,7	18,1	18,5
heinäkuu		15,2	18,6	17,8	18,2
elokuu		14,9	16,9	14,7	16,6
toukokuu	netti	20,4	24,0	36,1	42,0
kesäkuu		17,3	25,6	37,0	41,3
heinäkuu		18,8	26,9	42,3	41,5
elokuu		20,2	29,0	46,9	45,7

Taulukko 4.9: Vastaustapajakauma prosentteina koulutusasteittain

Yleisellä tasolla kannattaa toki kiinnittää huomiota siihen, että 2200 vastaajasta noin tuhat, eli noin 45 prosenttia, on keskiasteen koulutuksen käyneitä (kuva 4.13). Tästä seuraa, että heidän käyttäytymisensä vaikuttaa eniten saatuihin tuloksiin. Suurin huolen aihe on kuitenkin vain peruskoulutuksen saaneet vastaajat, joista vain noin kolmannes vastaa kyselyyn. Näistä vastaajista noin puolet vastaa netitse ja loput puhelimitse.

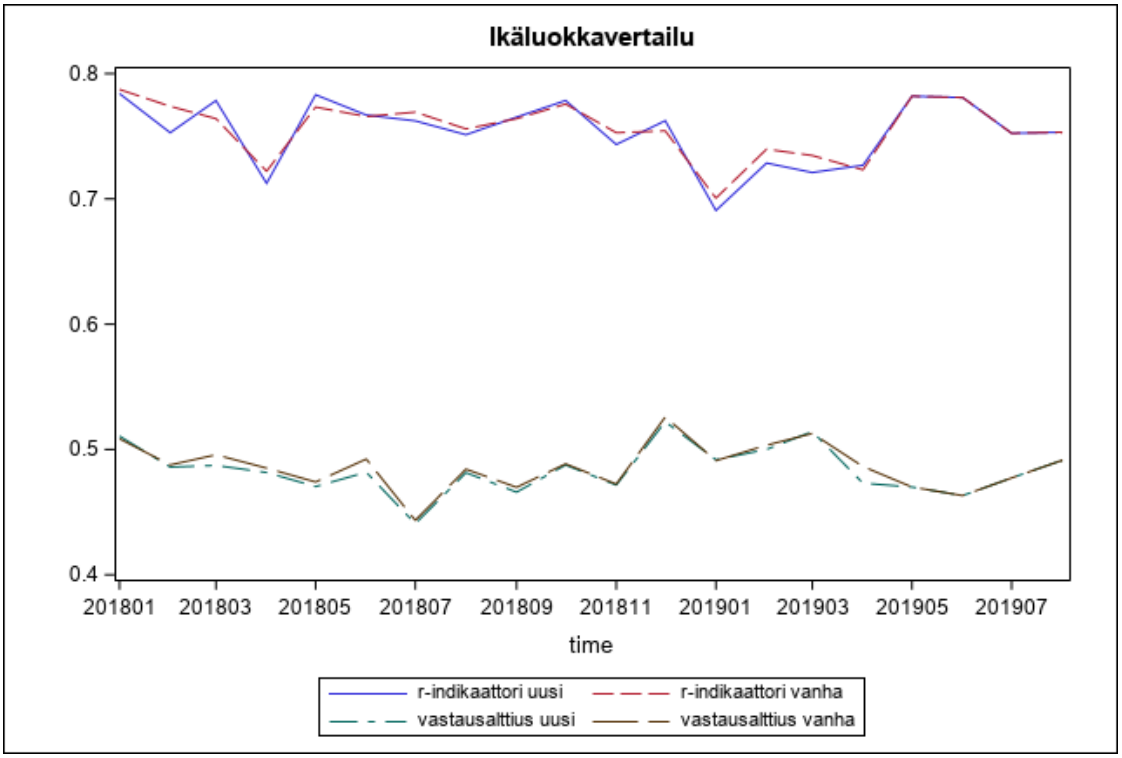


Kuva 4.13: Vastaustapa yhdistelmätiedonkeruun ajalta koulutusasteittain

Muut käytettävissä olevat taustamuuttujat (sukupuoli ja alue) eivät merkittävästi erottele vastaamistapoja, joten ne voidaan sivuuttaa, kuten luvussa 4.1.1 mainittiin.

4.2.2 R-indikaattori

Yhdistelmätiedonkeruuseen siirtymisen lisäksi myös kyselyyn osallistuneiden ikäjakaumaa muutettiin toukokuun alussa. Vanhin ikäluokka (75–84) poistettiin kokonaan ja nuorimmasta ikäluokasta (ennen 15–24) jätettiin alaikäiset pois (nykyään 18–24). Tämän muutoksen vaikutusta R-indikaattoriin on suhteellisen helppo tutkia. Kun vanhasta aineistosta jätetään pois mainitut ikäryhmät, voidaan laskea varsin vertailukelpoisia R-indikaattoreita. Kuvassa 4.14 ja taulukossa 4.10 nähdään, että mainittujen ikäryhmien pois jätöllä ei ole juuri minkäänlaista vaikutusta R-indikaattoreihin tai vastausasteisiin. R-indikaattorin arvo on laskenut keskimäärin $0,25$ prosenttiyksikköä erotuksen keskihajonnan ollessa $0,008$. Vastausaste puolestaan on laskenut keskimäärin $0,35$ prosenttiyksikköä erotuksen keskihajonnan ollessa $0,004$. Voidaan siis sanoa, että näillä mittareilla ikäjakauman supistamisella ei ole vaikutusta otoksen yleiseen edustavuuteen. Muutoksen seurauksena otoskoot hieman vaihtelevat, kun kyseisiin ikäluokkiin kuuluneet vastaajat on poistettu aineistosta.

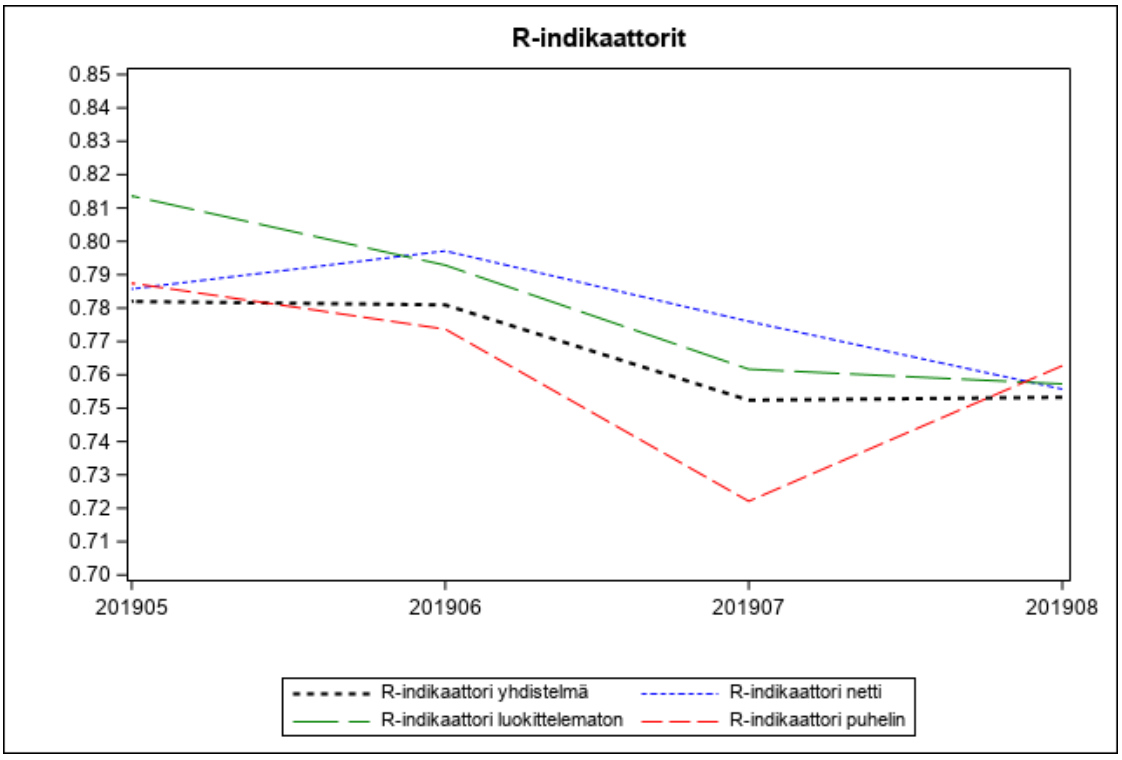


Kuva 4.14: Ikäluokkien muutoksen vaikutus vastausalttiuteen ja R-indikaattorin arvoihin

	R-indikaattori		**vastausaste**		**otoskoko**
kuukausi	barometri	luottamus	barometri	luottamus	luottamus
2018-05	0,773	0,783	0,474	0,470	2060
2018-06	0,766	0,767	0,492	0,482	2063
2018-07	0,769	0,762	0,443	0,441	2067
2018-08	0,756	0,751	0,484	0,481	2079
2018-09	0,764	0,765	0,470	0,466	2054
2018-10	0,776	0,779	0,489	0,488	2053
2018-11	0,753	0,743	0,472	0,472	2059
2018-12	0,754	0,762	0,526	0,522	2057
2019-01	0,701	0,691	0,491	0,492	2072
2019-02	0,740	0,729	0,503	0,500	2054
2019-03	0,735	0,721	0,513	0,514	2061
2019-04	0,723	0,727	0,486	0,473	2080
2019-05		0,782		0,470	2200
2019-06		0,781		0,463	2200
2019-07		0,752		0,477	2200
2019-08		0,753		0,491	2200

Taulukko 4.10: R-indikaattorit ja vastausasteet uudella (luottamus) ja vanhalla (barometri) ikäluokituksella

Yhdistelmätiedonkeruun ajalta laskettiin myös kuvitteelliset R-indikaattorit tilanteessa, jossa huomioidaan joko vain netti- tai puhelinvastaajat. Nämä laskelmat ovat toki virheellisiä, sillä ne sisältävät kaikki vastaamattomat riippumatta siitä kumpaa haastattelumenetelmään käytettiin. Tästä johtuvat taulukossa 4.11 näkyvät hyvin pienet vastausasteet. Todellisuudessahan on näyttöä siitä, että suuri osa nettivastaajista olisi vastannut puhelinhaastatteluun, jollei netissä vastaaminen olisi ollut mahdollista. Tämä laskenta tehtiin siis kokeellisesta mielenkiinnosta ja lasketuista tuloksista ei kannata vetää sen pidemmälle meneviä johtopäätöksiä aineiston suhteen.



Kuva 4.15: R-indikaattoreiden kuvitteelliset erot eri vastaustavoilla

Havaitaan, että nettivastaajat edustavat parhaiten perusjoukkoa, jos R-indikaattoria käytetään mittarina. Vastaavasti pelkkä puhelin on vähiten edustava samalla mittarilla. Mielenkiintoiselta vaikuttava yksityiskohta on, että yhdistelmätiedonkeruu saa huonommat tulokset kuin nettikysely.

Edellä mainittujen laskelmien lisäksi kuvaajassa 4.15 on myös R-indikaattorin arvot tilanteessa, jossa ikää ei käytetä luokiteltuna laatueroasteikollisena muuttujana, vaan sen annetaan olla suhdeasteikollinen. Havaitaan, että uudelleen luokiteltaessa ikä laatueroasteikolliseksi R-indikaattorin arvot laskevat hieman. Erot tuloksissa eivät ole merkittäviä, mutta tällaiseen luokitteluun ei ole mitään varsinaista tarvetta, jollei ole tarkoituksena laskea myös osittaisia R-indikaattoreita.

Yhdistelmätiedonkeruuseen siirtymisen vaikutuksesta R-indikaattoriin suhteessa aikaan ei kertyneen aineiston perusteella pääse juurikaan ottamaan kantaa. Luvussa 4.1.5 aineistoon sovitettiin lineaariset regressiomallit, joissa ajalla selitettiin R-indikaattorin ja vastausasteiden muutosta. Viimeisten neljän kuukauden aikana

R-indikaattori	yhdistelmä	netti	puhelin
toukokuu	0,782	0,786	0,787
kesäkuu	0,781	0,797	0,774
heinäkuu	0,752	0,776	0,722
elokuu	0,753	0,756	0,763
"vastausaste"	yhdistelmä	netti	puhelin
toukokuu	0,470	0,346	0,263
kesäkuu	0,463	0,345	0,252
heinäkuu	0,477	0,365	0,252
elokuu	0,491	0,395	0,239

Taulukko 4.11: R-indikaattorit ja niiden laskennassa käytetyt vastausasteet

R-indikaattorin arvot olivat toki ylempänä kuin lineaarisen regressiomallin antamat ennusteet, mutta neljän havainnon perusteella ei valitettavasti pääse tekemään suuria päätelmiä (taulukko 4.12). Ja mitä taas tulee vastausasteisiin, niin ne vaihtelivat lineaarisen regressiomallin ennusteiden molemmin puolin (taulukko 4.12).

	Vastausalttius	ennuste		R-indikaattori	ennuste
toukokuu	0,4700	0,4702		0,7733	0,7413
kesäkuu	0,4632	0,4685		0,7724	0,7407
heinäkuu	0,4773	0,4668		0,7448	0,7402
elokuu	0,4914	0,4651		0,7457	0,7397

Taulukko 4.12: Lineaarisen regressiomallin ennusteet yhdistelmätiedonkeruun ajalta

Luku 5

Johtopäätökset

5.1 Kadon ja edustavuuden kehitys

Taulokosta 5.1 nähdään, että vastausaste on pudonnut vuosien 2012 ja 2018 välillä yli 10 prosenttiyksikköä. Viimeisten puhelinkyselyyn perustuvien kuukausien (tammikuu–hutikuu 2019) vastausprosenttien keskiarvo oli hieman korkeampi kuin edellisen vuoden keskiarvo, mutta edelleen alle 50 prosenttia. Luonnollisestikaan tätä neljän kuukauden ajanjaksoa ei voi täysimääräisesti verrata muihin tuloksiin, koska vastausasteissa on vuoden sisäistä vaihtelua. Vastausasteiden lasku on kuitenkin selvä ja sille on syytä tehdä jotain. Ja näinhän Tilastokeskus on jo tehnytkin.

vuosi	2012	2013	2014	2015	2016	2017	2018		2019
R-indikaattori	0,774	0,769	0,758	0,756	0,752	0,746	0,753		0,710
vastausaste	0,618	0,606	0,547	0,546	0,516	0,507	0,482		0,495

Taulukko 5.1: R-indikaattorin ja vastausasteen muutos

Samasta taulukosta (5.1) nähdään että edustavuudessa ei ole tapahtunut vastaavaa laskua. R-indikaattori on laskenut kuudessa vuodessa vain noin 0.02 yksikköä. Tällä mittarilla mitattaessa saisi sen vaikutelman, että vastausasteen laskulla ei olisi suurta väliä, kunhan edustavuus vain säilyy. Ja näinhän se periaatteessa olisikin.

R-indikaattorille määriteltiin kuitenkin kolme hyvyysmittaria luvussa 3.2.8. Näistä hyvyysmittareista RR_1:n ja RR_3:n arvoja ei oikeastaan koskaan saavutettu. RR_2:n arvot sen sijaan olivat saavutettavissa, niin kauan kuin vastausasteet nousivat yli 60:n prosentin, mutta siitä on aikaa jo useita vuosia. Koska R-indikaattorien saamista yksittäisistä lukuarvoista ei voi tehdä suoria johtopäätöksiä, niin täytyy tukeutua näihin hyvyysrajoihin. Tällöin täytyy todeta, että saadut R-indikaattorien arvot ovat liian alhaisia koko mittauskaudella, eli syytä huoleen on. Toisaalta on muistettava, että R-indikaattorin arvo riippuu käytetyistä apumuuttujista. Tässä tutkimuksessa oli tarjolla neljä apumuuttujaa, joista kahta parasta käytettiin mallinnuksessa. Joillain toisilla apumuuttujilla tulos olisi voinut olla parempi.

5.2 Yhdistelmätiedonkeruu

Taulukosta 5.2 nähdään, että yhdistelmätiedonkeruuseen siirtyminen ei neljän ensimmäisen kuukauden (toukokuu–elokuu 2019) perusteella nostanut vastausalttiutta, vaan peräti laski sitä hieman. Toki kannattaa huomioida, että kyseessä on kesäkuukaudet, joiden vastausasteet olivat edellisinäkin vuosina hieman muuta vuotta alemmalla tasolla.

jakso	2017,I	2017,II	2017,III	2018,I	2018,II	2018,III	2019,I	2019,II
R-ind.	0,737	0,747	0,756	0,749	0,757	0,754	0,710	0,759
v-aste	0,515	0,501	0,504	0,491	0,469	0,487	0,495	0,475

Taulukko 5.2: R-indikaattorin ja vastausasteen muutos neljän kuukauden jaksoissa

Mitä taas tulee R-indikaattorin arvoihin, niin ne saivat tällä kahden vuoden jaksolla korkeimman arvonsa yhdistelmätiedonkeruukuukausina. Viimeksi yhtä korkeaan keskiarvoon on päästy loppuvuodesta 2015. Koska nämä keskiarvot ovat kuitenkin suunnilleen samalla tasolla kuin muinakin jaksoina, niin tästä ei voi lähteä vetämään mitään erityisiä johtopäätöksiä.

5.3 Vaihtoehtoiset indikaattorit

Osittaisista R-indikaattoreista pystyttiin päättelemään, että kaksi vastaajaryhmää on aliedustettuna lopullisessa aineistossa (kuva 4.7). Nämä ovat nuorimmat ikäluokat ja vain perusasteen koulutuksen saaneet. Tämä tieto on toki saatavissa myös luokitellun aineiston vastausasteiden kautta. Osittaisten R-indikaattorien hyödyntäminen otannan säätämisessä voisi olla hyvä idea, mutta se kaipaa lisätutkimusta.

Tämän tutkielman puitteissa havaittiin, että osa vaihtoehtoisista indikaattoreista (R_2, V, R_3, C_m) tuottaa lähes samoja tuloksia, kuin varsinainen R-indikaattori. Kovin eksaktisti tätä ei tutkittu, mutta lineaarimuunnosten jälkeen kuvaajien profiilit muistuttivat merkittävästi toisiaan (kuva 4.11). IMB-estimaattori ja CV poikkesivat hieman enemmän, mutta näilläkin perusprofiili oli hyvinkin samankaltainen. BIX, MAB ja Q_2 poikkesivat jo selkeämmin. R-indikaattorin ja vaihtoehtoisten indikaattorien korrelaatiokertoimet (taulukko 4.6) tukivat kuvaajaprofiileista katsottua tulkintaa. R-indikaattorin korvaaminen, jollain näistä vaihtoehtoisista edustavuusindikaattoreista voisi olla toimiva idea, jos jollain näistä indikaattoreista olisi selkeämpi tulkinta kuin R-indikaattorilla. Edustavuuden suhteen tällaista selkeää tulkintaa ei kuitenkaan ilmennyt.

5.4 Mitä voidaan tehdä?

Tämän tutkimuksen perusteella vaikuttaa siltä, että yhdistelmätiedonkeruuseen siirtyminen oli välttämätön, mutta ei riittävä askel taistelussa vastauskatoa vastaan. Jotta edustavuus saataisiin nousemaan, tulisi kaikki liikenevät voimavarat satsata heikoiten vastaaviin vastaajaryhmiin. Vaihtoehtoina on kasvattaa näiden

vastaajaryhmien osuuksia jo otantavaiheessa tai sitten yrittää nostaa heidän vastausalttiuttaan.

Viimeksi mainittu ei liene helppoa, koska vastaajia ei varsinaisesti voi painostaa vastaamaan. Kaupallisella puolella varmaankin palkkion maksaminen tulisi kyseeseen, mutta julkisella puolella tämä ei liene hyväksyttävä ratkaisu. Ja vaikka vastaajien motivointi palkkiolla tai painostamalla olisikin mahdollista, niin se ei kuitenkaan takaisi saatujen vastausten aitoutta. Vastaamisen pelillistäminen voisi houkutella nuoria ikäluokkia vastaamaan, mutta tämä idea ei varsinaisesti liity tämän tutkimuksen aihepiiriin.

Eli jäljelle jää heikosti vastaavien vastaajaryhmien (nuoret ikäluokat ja matalan koulutustason omaavat) vaikutuksen kasvattaminen otoksessa. Tämä ongelma on toki huomioitu tälläkin hetkellä, sillä lopullinen aineisto painotetaan. Painotuksessa käytetään kaikkia neljää taustamuuttujaa (ikä, koulutustaso, sukupuoli ja asuinalue), jotka olivat tässäkin tutkimuksessa esillä. Tämä painotus ei näy näissä tuloksissa, jotka on laskettu painottamattomilla aineistoilla. Painotuksen lisäksi voisi kokeilla myös näiden vastaajaryhmien osuuden aitoa kasvattamista otoksessa, jotta heidän kannoistaan saataisiin varmempi käsitys.

Toinen asia, joka jo saattaa korjata tilannetta, on siirtyminen rotatoivaan paneeliin, joka toteutettiin samalla, kun siirryttiin yhdistelmätiedonkeruuseen. Toki tällöin uusien vastaajien määrä kuukausittain pienenee, mutta vastaavasti muutokset ajassa tarkentuvat, kun samat henkilöt vastaavat kolmen kuukauden kuluttua uudestaan. Tämän muutoksen vaikutusta ei tietenkään pysty mitenkään mittaamaan käytössä olevan aineiston puitteissa, koska vasta aloitetun rotaation vaikutukset tulevat esiin vasta tulevina kuukausina.

Kirjallisuutta

[Bethlehem (1988)] Bethlehem, Jelke G. Reduction of Nonresponse Bias Through Regression Estimation (1988), Journal of Official Statistics, Vol. 4, No. 3, 1988, pp. 251–260

[Cornesse (2018)] Cornesse, C. & Bosnjak, M. (2018) Is there an association between survey characteristics and representativeness? A meta-analysis. Survey Research Methods Vol. 12, No. 1, pp. 1–13

[DG-ECFIN (2019)] European Comission Directorate-General for Economic and Financial Affairs. (2019) The Joint Harmonised EU Programme of Business and Consumer Surveys - User Guide,[viitattu: 17.10.2019]. Saantitapa: `https://ec.europa.eu/info/sites/info/files/bcs_user_guide_en_0.pdf`

[Dillman (2014)] Dillman, Don A., et al. (2014) Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. John Wiley & Sons, Incorporated.

[Groves (2009)] Groves, Robert M. (2009) Survey Methodology 2nd edition. John Wiley & Sons.

[Groves (2006)] Groves, Robert M. (2006) Nonresponse Bias in Household Surveys. The Public Opinion Quarterly Vol. 70, No. 5, Special Issue, pp. 646–675

[de Heij (2010)] de Heij, V., Schouten B. & Shlomo N. (2010) RISQ manual: Tools in SAS and R for the computation of R-indicators and partial R-indicators, RISQ deliverable 12.1

[Horvitz(1952)] Horvitz, Daniel G. & Thompson, Donovan J. (1952) A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, Vol. 47, No. 260 (1952), pp. 663–685

[Hosmer(2013)] Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. 2013. Applied logistic regression. 3rd ed. Hoboken, N.J.: Wiley.

[Laaksonen (2018)] Laaksonen, S. (2018) Survey Methodology and Missing Data. Springer.

[Laaksonen (2013)] Laaksonen, S. & Heiskanen, M. (2013) Comparison of three modes for a crime victimization survey. Journal of Survey Statistics and Methodology (2014) 2, 459–483

[de Leeuw (2005)] de Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 233–255

[Lehmann (1951)] Lehmann, E. L. (1951) A General Concept of Unbiasedness. The Annals of Mathematical Statistics, vol. 22, no. 4 , 1951, pp. 587–592.

[Little (2002)] Little, R.J.A. & Rubin, D. (2002) Statistical Analysis with Missing Data 2nd edition. John Wiley & Sons.

[Moilanen (2011)] Moilanen, Raija (2011) Barometrit. Kielikello Vol. 1, (2011).

[Nagelkerke (1991)] Nagelkerke, N.J.D. (1991) Miscellanea, A note on a general definition of the coefficient of determination. Biometrika, 78(3) pp. 691–692.

[Plewes (2013)] Plewes, T.J. & Tourangeau, R. (2013) Nonresponse in Social Science Surveys: A Research Agenda. National Academies Press.

[RISQ 2007] RISQ – Representative Indicators for Survey Quality (2007),[viitattu: 13.3.2020]. Saantitapa: `https://www.cmi.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/`

[Rosen (1970)] Rosen, S. & Tesser, A. (1970) On Reluctance to Communicate Undesirable Information: The MUM Effect. Sociometry Vol. 33, No. 3 (1970), pp. 253–263.

[Rosenbaum (1983)] Rosenbaum, Paul R. & Rubin, Donald B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, Vol. 70, No. 1, pp. 41–55.

[Rubin (1976)] Rubin, D. B. (1976). Inference and Missing Data. Biometrika, Vol. 63, No. 3, pp. 581–592.

[Schouten (2009a)] Schouten, B. & Bethlehem, J. (2009). Representativeness indicators for measuring and enhancing the composition of survey response.

[Schouten (2007)] Schouten, B. & Cobben, F.: R-indexes for the comparison of different fieldword strategies and data collection modes. Discussion paper 07002, Voorburg/Heerlen (2007).

[Schouten (2009b)] Schouten, B., Cobben, F. & Bethlehem, J.: Indicators for the representativeness of survey response *Survey Methodology*, Vol. 35, No. 1, pp. 101–113, Statistics Canada (2009).

[Shlomo (2009a)] Shlomo, N., Skinner, C.J., Schouten, B., Bethlehem, J. & Zhang, Li-Chun (2009). Statistical Properties of R-indicators.

[Shlomo (2009b)] Shlomo, N., Skinner, C.J., Schouten, B., Carolina, T. & Morren, M. (2009). Partial Indicators for Representative Response.

[Shlomo (2012)] Shlomo, N., Skinner, C.J. & Schouten, B. (2012). Estimation of an Indicator of the Representativeness of Survey Response. Journal of Statistical Planning and Inference 142, pp. 201–211.

[Shlomo (2013)] Shlomo, N. & Schouten, B. (2013) Theoretical Properties of Partial Indicators for Representative Response.

[SVT (2019)] Suomen virallinen tilasto (SVT): Kuluttajabarometri [verkkojulkaisu]. ISSN=1796-864X. Huhtikuu 2019, Laatuseloste: Kuluttajabarometri . Helsinki: Tilastokeskus [viitattu: 6.5.2019]. Saantitapa: http://www.stat.fi/til/kbar/2019/04/kbar_2019_04_2019-04-29_laa_001_fi.html

[SVT (2017)] Suomen virallinen tilasto (SVT): Kuluttajien luottamus [verkkojulkaisu]. ISSN=2669-8862. Helsinki: Tilastokeskus [viitattu: 17.10.2019]. Saantitapa: http://www.stat.fi/til/kbar/kbar_2017-05-05_men_001.html

[SVT (2019)] Suomen virallinen tilasto (SVT): Kuluttajien luottamus [verkkojulkaisu]. ISSN=2669-8862. Syyskuu 2019, Laatuseloste: Kuluttajien luottamus . Helsinki: Tilastokeskus [viitattu: 17.10.2019]. Saantitapa: http://www.stat.fi/til/kbar/2019/09/kbar_2019_09_2019-09-27_laa_001_fi.html

[Särndal (2016)] Särndal, C.-E., Lumiste, K. & Traat, I. (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 42, No. 2.

[Särndal (2005)] Särndal, C.-E. & Lundström S. (2005). Estimation in Surveys with Nonresponse. John Wiley & Sons.

[Zhang (2013)] Zhang, L., Thomsen I. & Öyvin, K. (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. International Statistical Review (2013), Vol. 81, No. 2.

Luku 6

Liitteet

6.1 SAS-koodi ja tulosteet

6.1.1 Tutkielmaa varten laadittu SAS-koodi

```
LIBNAME UUS 'D:\UD\hyhko\Documents\My SAS Files\kbar\uudet';

data UUS.mabit;
set UUS.kulu201907;
mabclass=ikal*10+tutk;
run;

proc logistic data=UUS.mabit NOPRINT;
 class ikal tutk;
    model respons= ikal tutk;
 score out=UUS.mabit;
run;

%macro HMM1;
proc sql noprint;
%do i = 1 %to 7;
 %do j = 1 %to 7 %by 2;
  select count(mabclass) into :MC&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select count(paino) into :MW&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select mean(P_1) format=best12. into :P_1&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select mean(paino1) format=best12. into :W_1&i&j from UUS.mabit
  where mabclass =&i*10+&j;
 %end;
%end;
quit;
```

```
data UUS.HMM;
set UUS.HMM;
%do k = 1 %to 7;
 %do l = 1 %to 7 %by 2;
  otos&k&l=symget('MC'||LEFT(&k&l));
  prop&k&l=symget('P_1'||LEFT(&k&l));
  vast&k&l=symget('MW'||LEFT(&k&l));
  pain&k&l=symget('W_1'||LEFT(&k&l));
 %end;
%end;
run;
%mend;

proc sql noprint;
select count(*) into :Otos from UUS.mabit;
select count(paino) into :Vast from UUS.mabit;
select sum(P_1*paino1) format=best12. into :Prop from UUS.mabit;
select sum(paino1) format=best12. into :Popu from UUS.mabit;
quit;

data UUS.mabit;
set UUS.mabit;
prop1=&Prop/&Popu;
r_i1=paino1*(P_1-prop1)**2;
run;

proc sql noprint;
select sum(r_i1) format=best12. into :Rind from UUS.mabit;
quit;

data UUS.HMM;
vast1=&Vast;
otos1=&Otos;
rate1=vast1/otos1;
popu1=&Popu;
prop1=&Prop/&Popu;
spro1=(&Rind/(&Popu-1))**(1/2);
Rindicator=1-2*spro1;
CV=spro1/prop1;
R_2=1-4*(&Rind/(&Popu-1));
%HMM1
run;

%macro HMM2;
data UUS.HMM;
set UUS.HMM;
MAB1=0;
IMB1=0;
Q1=0;
```

```
N1=0;
X21=0;
L1=0;
%do m = 1 %to 7;
 %do n = 1 %to 7 %by 2;
  if otos&m&n>0 then
  do;
   N1=N1+1;
   AB1&m&n=ABS(vast&m&n/vast1-otos&m&n/otos1);
   IMB1=IMB1+(otos&m&n/otos1)*(vast&m&n/otos&m&n-vast1/otos1)**2;
   if vast&m&n>0 then Q1=Q1+(otos&m&n**2/(vast&m&n*vast1));
   X21=X21+otos&m&n*(prop&m&n-prop1)**2;
   L1=L1+LOG(Prop&m&n**(Otos&m&n*Prop&m&n)*
        (1-Prop&m&n)**((otos&m&n*(1-Prop&m&n))));
  end;
 %end;
%end;

L01=(prop1**(2*prop1))*(1-prop1)**(2*(1-prop1));
L11=EXP(L1*2/otos1);
R_3=1-(1-L01/L11)/(1-L01);
V=(X21/(prop1*(1-prop1)*(otos1)))**(1/2);
Q2=Q1-otos1**2/vast1**2;
IMB=IMB1;
BIX=sum(of AB1:);
MAB=median(of AB1:);
MinAB=min(of AB1:);
MeanAB=mean(of AB1:);
MaxAB=max(of AB1:);
rate=rate1;
prop=prop1;
Bm=(1-Rindicator)/(2*prop1);
Cm=(1-Rindicator)/(2*prop1*(1-prop1));
gamma=0.05;
RR1_05=1-gamma*2/quantile('NORMAL',0.975,0,1);
RR2_05=1-2*prop1*gamma;
RR3_05=1-2*prop1*(1-prop1)*gamma;
gamma=0.1;
RR1_10=1-gamma*2/quantile('NORMAL',0.975,0,1);
RR2_10=1-2*prop1*gamma;
RR3_10=1-2*prop1*(1-prop1)*gamma;
gamma=0.2;
RR1_20=1-gamma*2/quantile('NORMAL',0.975,0,1);
RR2_20=1-2*prop1*gamma;
RR3_20=1-2*prop1*(1-prop1)*gamma;
uusi=0;
vast=vast1;
otos=otos1;
luokkia=N1;
```

```
run;
%mend;

%HMM2;
run;
/*
proc export
  data=UUS.hmm
  dbms=xlsx
  outfile="D:\UD\hyhko\Documents\My SAS Files\kbar\uudet\hmm.xlsx"
  replace;
run;
*/

/*Uusi pienempi otos (18—74)*/

data UUS.mabit;
set UUS.mabit;
if ika < 18 or ika >74 then delete;
run;

proc logistic data=UUS.mabit NOPRINT;
 class tutk;
 model respons= ikal tutk;
 score out=UUS.mabit;
run;

%macro HMMM1;
proc sql noprint;
%do i = 1 %to 6;
 %do j = 1 %to 7 %by 2;
  select count(mabclass) into :MC&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select mean(P_12) format=best12. into :P_12&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select sum(respons2) into :MW&i&j from UUS.mabit
  where mabclass =&i*10+&j;
  select mean(paino2) format=best12. into :W_12&i&j from UUS.mabit
  where mabclass =&i*10+&j;
 %end;
%end;
quit;

data UUS.HMMM;
set UUS.HMMM;
%do k = 1 %to 6;
 %do l = 1 %to 7 %by 2;
  otos&k&l=symget('MC'||LEFT(&k&l));
  prop&k&l=symget('P_12'||LEFT(&k&l));
```

```
  vast&k&l=symget('MW'||LEFT(&k&l));
  pain&k&l=symget('W_12'||LEFT(&k&l));
 %end;
%end;
run;
%mend;

proc sql noprint;
select count(*) into :Otos from UUS.mabit
where ika > 17 and ika < 75;
select sum(respons2) into :Vast from UUS.mabit
where ika > 17 and ika < 75;
select sum(P_12*paino2) format=best12. into :Prop from UUS.mabit
where ika > 17 and ika < 75;
select sum(paino2) format=best12. into :Popu from UUS.mabit
where ika > 17 and ika < 75;
quit;

data UUS.mabit;
set UUS.mabit;
prop2=&Prop/&Popu;
r_i2=paino2*(P_12-prop2)**2;
run;

proc sql noprint;
select sum(r_i2) format=best12. into :Rind from UUS.mabit;
quit;

data UUS.HMMM;
vast2=&Vast;
otos2=&Otos;
rate2=vast2/otos2;
popu2=&Popu;
prop2=&Prop/&Popu;
spro2=(&Rind/(&Popu-1))**(1/2);
Rindicator=1-2*spro2;
CV=spro2/prop2;
R_2=1-4*(&Rind/(&Popu-1));
%HMMM1
run;

%macro HMMM2;
data UUS.HMMM;
set UUS.HMMM;
MAB2=0;
IMB2=0;
Q22=0;
N2=0;
X22=0;
```

```
L2=0;
%do m = 1 %to 6;
 %do n = 1 %to 7 %by 2;
  if otos&m&n>0 then
  do;
   N2=N2+1;
   AB2&m&n=ABS(vast&m&n/vast2-otos&m&n/otos2);
   IMB2=IMB2+(otos&m&n/otos2)*(vast&m&n/otos&m&n-vast2/otos2)**2;
   if vast&m&n>0 then Q22=Q22+(otos&m&n**2/(vast&m&n*vast2));
   X22=X22+otos&m&n*(prop&m&n-prop2)**2;
   L2=L2+LOG(Prop&m&n**(Otos&m&n*Prop&m&n)*
        (1-Prop&m&n)**((otos&m&n*(1-Prop&m&n))));
  end;
 %end;
%end;

L02=(prop2**(2*prop2))*(1-prop2)**(2*(1-prop2));
L12=EXP(L2*2/otos2);
R_3=1-(1-L02/L12)/(1-L02);
V=(X22/(prop2*(1-prop2)*(otos2)))**(1/2);
Q2=Q22-otos2**2/vast2**2;
IMB=IMB2;
BIX=sum(of AB2:);
MAB=median(of AB2:);
MinAB=min(of AB2:);
MeanAB=mean(of AB2:);
MaxAB=max(of AB2:);
rate=rate2;
prop=prop2;
Bm=(1-Rindicator)/(2*prop2);
Cm=(1-Rindicator)/(2*prop2*(1-prop2));
gamma=0.05;
RR1_05=1-2*gamma/quantile('NORMAL',0.975,0,1);
RR2_05=1-2*prop2*gamma;
RR3_05=1-2*prop2*(1-prop2)*gamma;
gamma=0.1;
RR1_10=1-2*gamma/quantile('NORMAL',0.975,0,1);
RR2_10=1-2*prop2*gamma;
RR3_10=1-2*prop2*(1-prop2)*gamma;
gamma=0.2;
RR1_20=1-2*gamma/quantile('NORMAL',0.975,0,1);
RR2_20=1-2*prop2*gamma;
RR3_20=1-2*prop2*(1-prop2)*gamma;
uusi=1;
vast=vast2;
otos=otos2;
luokkia=N2;
run;
%mend;
```

```
%HMMM2;
run;

data UUS.MAB201907;
set UUS.HMM
UUS.HMMM;
month=201907;
month_char = PUT(month, 6.0);
time = INPUT(month_char, yymmn6.);
keep vast otos rate prop MAB MeanAB BIX Rindicator R_2 R_3
   CV IMB Q2 V Bm Cm RR1_05 RR1_10 RR1_20 RR2_05 RR2_10 RR2_20
   RR3_05 RR3_10 RR3_20 luokkia month time uusi;
run;
```

6.1.2 Tutkielmaa varten laaditun SAS-koodin tuloste, 7/2019

Huom: Alla olevasta tulosteesta on jätetty pois *R*-indikaattorien hyvyysrajojen isompien γ:n arvot (RR1_10 jne.), jotka on myös laskettu liitteessä 6.1.1 olevalla SAS-koodilla.

Rindicator	CV	R_2	R_3	V
0.7448479577	0.2673021386	0.9348974353	0.9925131409	0.2554160059

Q2	IMB	BIX	MAB	MeanAB
0.5039841506	0.0176710359	0.2181818182	0.009004329	0.009486166

rate	prop	Bm	Cm
0.4772727273	0.4772727289	0.2673021386	0.5113606147

RR1	RR2	RR3	uusi
0.9489786543	0.9522727271	0.9750516529	0

vast	otos	luokkia	month	time
1050	2200	23	201907	21731

6.1.3 RISQ-projektin SAS-koodin tuloste, 7/2019

Huom: Alla olevasta tulosteesta on jätetty pois osittaiset *R*-indikaattorit, koska näiden taulukoiden luettavuus olisi merkittävästi kärsinyt tarvittavasta formaatin muutoksesta.

r_indicator	r_withbias	avgprop	LB_r
0.7523828391	0.7448479577	0.4772727289	0.7129565209

UB_r	variance_prop	variance_prop_adj	StdErr_r
0.7918091573	0.0162756412	0.0153285646	0.0201154685

CV_prop_adj	StdErr_CV_adj	CV_prop_unadj	StdErr_CV_unadj
0.2673021387	0.0211283345	0.2594084534	0.0211221286

6.2 Kuluttajabarometrin kysymykset

6.2.1 EU-harmonisoidut kysymykset

1. Mikä on kotitaloutenne taloustilanne nyt verrattuna tilanteeseen 12 kuukautta sitten?
2. Mikä on kotitaloutenne taloustilanne 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
3. Mikä on Suomen taloudellinen tila nyt verrattuna tilanteeseen 12 kuukautta sitten?
4. Mikä on Suomen taloudellinen tila 12 kuukauden kuluttua verrattuna tilanteeseen nyt?
5. Millä tasolla kuluttajahinnat ovat nyt verrattuna tilanteeseen 12 kuukautta sitten?
6. Montako prosenttia kuluttajahinnat ovat muuttuneet viimeisen 12 kuukauden aikana?
7. Miten kuluttajahinnat muuttuvat seuraavan 12 kuukauden aikana?
8. Montako prosenttia kuluttajahinnat muuttuvat seuraavan 12 kuukauden aikana?
9. Miten työttömyystilanne muuttuu Suomessa seuraavan 12 kuukauden aikana?
10. Onko nyt yleisesti ottaen hyvä vai huono aika hankkia kestokulutustavaroita?
11. Käytättekö kestokulutustavaroiden hankintaan rahaa seuraavan 12 kuukauden aikana enemmän, yhtä paljon vai vähemmän kuin viimeisen 12 kuukauden aikana?
12. Onko nyt yleisesti ottaen hyvä vai huono aika säästää?
13. Kuinka todennäköisesti kotitaloutenne pystyy säästämään rahaa seuraavan 12 kuukauden aikana?
14. Mikä on kotitaloutenne rahatilanne tällä hetkellä?
15. Kuinka todennäköisesti kotitaloutenne ostaa henkilöauton 12 kuukauden sisällä?
16. Aikooko kotitaloutenne ostaa tai rakentaa asunnon 12 kuukauden sisällä?
17. Kuinka todennäköisesti kotitaloutenne käyttää suuren summan rahaa kodin perusparannuksiin 12 kuukauden sisällä?

6.2.2 Tilastokeskuksen omat barometrikysymykset

1. Aikooko kotitaloutenne hankkia uuden vai käytetyn auton?
2. Miten aiotte rahoittaa autonne oston (2 tärkeintä rahoitusmuotoa)? (kysytään neljästi vuodessa)
3. Miten aiotte rahoittaa asuntonne oston (2 tärkeintä rahoitusmuotoa)? (kysytään neljästi vuodessa)

4. Aikooko kotitaloutenne käyttää rahaa seuraaviin hyödykeryhmiin 6 kuukauden sisällä: asunnon korjaus, kodin sisustus, loma-asunto, viihde-elektroniikka, kodinkoneet, kulkuvälineet (ei auto), harrastus- ja urheiluvälineet, lomamatka kotimaassa, lomamatka ulkomaille?

5. Mitä tarkoitusta varten aiotte säästää? (kysytään neljästi vuodessa)

6. Mihin kohteisiin aiotte sijoittaa säästöjänne? (kysytään neljästi vuodessa)

7. Onko nyt yleisesti ottaen hyvä vai huono aika ottaa lainaa?

8. Aikooko kotitaloutenne ottaa lainaa 12 kuukauden sisällä?

9. Mitä tarkoitusta varten aiotte ottaa lainaa? (kysytään neljästi vuodessa)

10. Kuinka arvioitte uhan jäädä itse työttömäksi muuttuneen viimeisen 12 kuukauden aikana?

11. Mitkä seuraavista laitteista kotitaloudellanne on (n. 25 laitetta: viihde-elektroniikka, tietotekniikka, puhelimet, auto)? (kysytään neljästi vuodessa)

6.2.3 Luokittelukysymykset

1. Montako henkilöä kotitalouteenne kuuluu?

2. Montako aikuista/lasta (4 ikäryhmää) kotitaloudessanne on?

3. Moniko kotitaloutenne jäsenistä käy säännöllisesti työssä?

4. Mikä on nykyinen kotikuntanne?

5. Mikä on kotitaloutenne asumismuoto?

6. Mikä on pääasiallinen toimintanne tällä hetkellä?

7. Mikä on ammattinne?

8. Onko teillä ammatillista koulutusta (nykyiseen) ammattiinne?

9. Mitkä ovat kotitaloutenne bruttotulot?

6.2.4 Taustatiedot

1. ikä

2. sukupuoli

3. otoskunta

4. koulutus

5. asuinkunta

6.3 Kuluttajien luottamus -tutkimuksen kysymykset

B1 Millainen on mielestäsi oma taloudellinen tilanteesi nyt, verrattuna tilanteeseen 12 kuukautta sitten?

- Paljon parempi, Jonkin verran parempi, Samanlainen, Jonkin verran huonompi, Paljon huonompi, EOS

B2 Entä millaisen arvioit sen olevan 12 kuukauden kuluttua, verrattuna tilanteeseen nyt?

- Paljon parempi, Jonkin verran parempi, Samanlainen, Jonkin verran huonompi, Paljon huonompi, EOS

B3 Millainen on sinun mielestäsi Suomen taloudellinen tilanne nyt, verrattuna tilanteeseen 12 kuukautta sitten?

- Paljon parempi, Jonkin verran parempi, Samanlainen, Jonkin verran huonompi, Paljon huonompi, EOS

B4 Entä millaisen arvioit sen olevan 12 kuukauden kuluttua, verrattuna tilanteeseen nyt?

- Paljon parempi, Jonkin verran parempi, Samanlainen, Jonkin verran huonompi, Paljon huonompi, EOS

B5 Kuinka paljon arvioit hintojen [nousseen / laskeneen] prosentteina viimeisen 12 kuukauden aikana? Voit antaa luvun yhden desimaalin tarkkuudella.

B6 Kuinka paljon arvioit hintojen [nousevan / laskevan] prosentteina seuraavan 12 kuukauden aikana? Voit antaa luvun yhden desimaalin tarkkuudella.

B7 Miten arvioit työttömien määrän muuttuvan Suomessa? Arveletko, että työttömiä on 12 kuukauden päästä:

B8 Onko työttömyyden tai lomautuksen uhka omalla kohdallasi viimeisen 12 kuukauden aikana mielestäsi:

C1 Jos ajattelet yleistä taloudellista tilannetta Suomessa, niin millainen aika mielestäsi nyt on ostaa kestokulutustavaroita, kuten huonekaluja, kodintekniikkaa tai auto?

C2 Jos ajattelet yleistä taloudellista tilannetta Suomessa, niin millainen aika mielestäsi nyt on säästää?

C3 Jos ajattelet taas yleistä taloudellista tilannetta, niin millainen aika mielestäsi nyt on ottaa lainaa?

D1 Mikä seuraavista vaihtoehdoista kuvaa parhaiten omaa rahatilannettasi tällä hetkellä?

D2 Kuinka todennäköistä on, että säästät rahaa seuraavan 12 kuukauden aikana? Myös velan lyhentäminen on säästämistä.

D3 Aiotko ottaa lainaa seuraavan 12 kuukauden aikana?

E1 Verrattuna edelliseen 12 kuukauteen, miten aiot käyttää rahaa kestokulutustavaroiden hankintaan seuraavan 12 kuukauden aikana?

E2 Kuinka todennäköistä on, että käytät rahaa henkilöauton ostoon seuraavan 12 kuukauden aikana?

E3 Aiotko käyttää rahaa asunnon ostoon tai talon rakentamiseen seuraavan 12 kuukauden aikana? Omaan tai jonkun perheenjäsenen käyttöön, loma-asunnoksi, vuokrattavaksi, myös rakenteilla oleva talo.

E4 Kuinka todennäköistä on, että käytät suuren summan rahaa asunnon korjauksiin tai parannuksiin seuraavan 12 kuukauden aikana? Esimerkiksi lämmitysjärjestelmän korjaus tai uusiminen, kylpyhuoneremontti, lattiarakenteiden uusiminen, rakennuksen laajentaminen.